

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FARMÁCIA
(FISIOPATOLOGIA E TOXICOLOGIA)
ÁREA DE TOXICOLOGIA

Raisa Ludmila Calil

Uma análise quimioinformática sobre canabinoides sintéticos
em receptores canabinoides (CB1 e CB2)

São Paulo

2023

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FARMÁCIA
(FISIOPATOLOGIA E TOXICOLOGIA)
ÁREA DE TOXICOLOGIA

Raisa Ludmila Calil

Uma análise quimioinformática sobre canabinoides sintéticos
em receptores canabinoides (CB1 e CB2)

Versão Original

Dissertação para obtenção de título de mestre da Faculdade de Ciências
Farmacêuticas da Universidade de São Paulo, área de Toxicologia
Orientador: Prof. Dr. Gustavo Henrique Goulart Trossini

São Paulo
2023

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

FICHA CATALOGRÁFICA

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

Bibliotecária responsável pela orientação de catalogação da publicação:
Marlene Aparecida Vieira - CRB - 8/5562

C153a	Calil, Raisa Ludmila Uma análise quimioinformática sobre canabinóides sintéticos em receptores canabinóides (CB1 e CB2) / Raisa Ludmila Calil. - São Paulo, 2023. 97 p. Dissertação (mestrado) - Faculdade de Ciências Farmacêuticas da Universidade de São Paulo. Departamento de Análises Clínicas e Toxicológicas. Orientador: Trossini, Gustavo Henrique Goulart 1. Canabinóides sintéticos. 2. Quimioinformática. 3. Random Forest. 4. MARS . 5. HCA e PCA. I. T. II. Trossini, Gustavo Henrique Goulart, orientador.
-------	--

Raisa Ludmila Calil

Uma análise quimioinformática sobre canabinoides sintéticos em receptores canabinoides (CB1 e CB2): dissertação para obtenção título de mestre da Faculdade de Ciências Farmacêuticas da Universidade de São Paulo, área de Toxicologia

Comissão Julgadora

Prof. Dr. orientador/presidente

1o. Examinador

2o. Examinador

3o. Examinador

4o. Examinador

São Paulo _____, _____ de 2023

Dedico esta dissertação aos meus pais,
Calil e Marly.

Agradeço à Universidade de São Paulo, à Faculdade de Ciências Farmacêuticas da USP pela oportunidade concedida. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Primeiramente, gostaria de expressar meu profundo agradecimento ao meu orientador, Dr. Gustavo, pela oportunidade proporcionada. Ele que acreditou em mim como pesquisadora e confiou tarefas que estavam além do meu escopo e conhecimento. Foi um desafio imenso e, de certa forma, uma verdadeira aventura! Muito obrigada. Agradeço pela paciência, preocupação, amizade, discussões estimulantes, conversas inspiradoras, cafés compartilhados e, acima de tudo, pelas palavras amigas nos momentos de desespero. Sou imensamente grata por ter sido adotada pelo senhor e por me tornar uma "*Litecquinha*".

Gostaria de expressar meu muito obrigada ao meu irmão Yuri e aos meus amigos Gustavo Marques e Artur Caminero. Ao longo desta jornada de trabalho, suportaram dia após dia, e me ensinaram tudo o que sei sobre a linguagem R e códigos, sendo fundamentais para a elaboração desta dissertação. Obrigada por toda a paciência e ajuda incansável que me ofereceram. Com toda certeza, esse trabalho não teria saído do planejamento sem vocês.

Agradeço aos meus pais, Calil e Marly, ao meu irmão Yasser, às minhas cunhadas e àqueles que acham que sou "titia-beakman". Agradeço todo o incentivo, apoio e pelas palavras encorajadoras como aquele "vamos trabalhar, titia".

Agradeço ao João e sua família, por todo suporte e por me receberem de braços abertos, estando sempre dispostos a me ajudarem.

Gostaria de expressar meu agradecimento aos meus amados *Litecquinhos*: Adenilson Lima, André Berndt, Carolina Rocha, Haifa Hassanie, Marissa El Hajje, Rafael Bello, Thabata Vinha, e Witor Ferraz. Sou imensamente grata por ter essa equipe maravilhosa ao meu lado, sempre dispostos a ajudar, com palavras amigas e ideias geniais que auxiliaram no desenvolvimento deste trabalho. Admiro profundamente cada membro desse grupo de pessoas extremamente inteligentes.

E, por último, minha gratidão aos *old-Litecquinhos* (André Novaes, Elys Lima, Evelin Fornari, Fernando Gatti, Glaucio Monteiro, João Anhaia, Renan Augusto, Thuane Duarte), bem como os colegas do Bloco 13 (Karine Prado e Mariana Darbem) que ao longo desses anos, eles foram fundamentais para moldar a pesquisadora que sou hoje. Agradeço por todas as discussões enriquecedoras, ajudas indispensáveis, carinho constante, preocupação genuína, momentos de descontração nos corredores e churrascos animados.

“Que você seja um girassol, que brota nessa rua de desatenção,
dessa vida vulgarmente cotidiana”
(Marsili, I.)

RESUMO

CALIL, R. L. **Uma análise quimiinformática sobre canabinoides sintéticos em receptores canabinoides (CB1 e CB2)**. 2023. 97. Dissertação (Mestrado) - Faculdade de Ciências Farmacêuticas, Universidade de São Paulo. 2023

Estudos científicos mostram o uso terapêutico da *Cannabis sativa* e de seus constituintes químicos, tais como: canabidiol, canabinol e tetrahydrocanabinol (THC), sendo este último responsável pelos efeitos psicoativos. Os receptores canabinoides CB1R e CB2R foram identificados, principalmente, no sistema nervoso central e imunológico, respectivamente. Canabinoides sintéticos são compostos que atuam nesses receptores, produzindo efeitos semelhantes à derivados canabinoides naturais. Técnicas computacionais são amplamente empregadas para avaliar seletividade molecular. Com o aumento das bibliotecas moleculares, se faz necessário explorar de forma racional as características estruturas e propriedades destes compostos. Neste estudo, utilizou-se técnicas de quimiinformática para gerar modelos que pudessem reconhecer e distinguir os perfis estruturais de canabinoides sintéticos e sua interação com os receptores canabinoides CB1R e CB2R. Para isso, foram realizadas: a construção de um banco de dados contendo informações sobre canabinoides sintéticos, cálculo de descritores moleculares, e construção de modelos utilizando técnicas como Análise Hierárquica de Agrupamentos (HCA), Análise de Componentes Principais (PCA), Árvores de Decisão, *Random Forest* e *Multivariate Adaptive Regression Splines (MARS)*. A seleção criteriosa dos descritores moleculares foi essencial para obter modelos estatísticos precisos. Quando se tratou de identificar as estruturas dos canabinoides sintéticos, o modelo baseado em árvore de decisão apresentou excelente desempenho na distinção da interação entre CB1R e CB2R. A redução da multicolinearidade mostrou que as abordagens de *Forward Selection* e *Backward Elimination* são eficazes na obtenção de modelos lineares simples e interpretáveis. O algoritmo *Random Forest* foi computacionalmente eficiente e proporcionou resultados confiáveis, enquanto o algoritmo *MARS* deve ser utilizado com cautela na predição de canabinoides sintéticos atuando em CB1R e CB2R. As técnicas de Análise do Componente Principal (PCA) e Análise Hierárquica de Agrupamentos (HCA) permitiram uma separação precisa dos dados e podem ser úteis em análises futuras.

Palavras- chave: Canabinoides sintéticos; Quimiinformática, *Random Forest*, *MARS*, *PCA*, *HCA*

ABSTRACT

CALIL, R. L. **Uma análise quimiinformática sobre canabinoides sintéticos em receptores canabinoides (CB1 e CB2)**. 2023. 97. Dissertação (Mestrado) - Faculdade de Ciências Farmacêuticas, Universidade de São Paulo. 2023

Tetrahydrocannabinol (THC), the chemical component that gives cannabis its euphoric effects, is present in *Cannabis sativa* and has been shown in scientific research to be used therapeutically. CB1R and CB2R cannabinoid receptors have been found primarily in the central nervous and immune systems, respectively. Synthetic cannabinoids are substances that bind to these receptors and result in cannabis-like effects. With the increase of molecular libraries, computational tools are helpful for finding specific molecules, thus it is necessary to explore rationally. In this study, chemoinformatics techniques were used to create models that could recognize and distinguish the structural profiles of synthetic cannabinoids and study their interaction with the cannabinoid receptors CB1R and CB2R. In order to achieve this, the following tasks were completed: building a database containing information on synthetic cannabinoids, estimating molecular descriptors, and constructing models using techniques like Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA), Decision Trees, Random Forest and Multivariate Adaptive Regression Splines (MARS). To create correct statistical models, it was crucial to choose chemical descriptors carefully. When it came to identifying the structures of synthetic cannabinoids that interact with CB1R and CB2R, the decision tree-based model performed exceptionally well. The decrease in multicollinearity demonstrated the efficacy of the Forward Selection and Backward Elimination procedures in producing straightforward and understandable linear models, while MARS algorithm should be used with caution in predicting synthetic cannabinoids acting on CB1R and CB2R. In contrast, the Random Forest approach was computationally efficient and provided reliable results. The data may be separated precisely thanks to the Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) approaches, which may be helpful in future research.

Keywords: Synthetic cannabinoids; Chemoinformatics, *Random Forest*, *MARS*, *PCA*, *HCA*

SUMÁRIO

1. Introdução	11
1.1 Canabinoides e seus derivados sintéticos	11
1.2 Quimioinformática	14
1.2.1 Descritores	14
1.2.2 <i>Machine Learning</i>	16
1.2.2.1 Regressão Linear	17
1.2.2.2 Árvores de Decisão	19
1.2.2.3 <i>Random Forest</i>	20
1.2.3.4 <i>MARS - Multivariate Adaptive Regression Splines</i>	22
1.2.2.5 <i>HCA - Hierarchical Cluster Analysis</i>	23
1.2.2.6 <i>PCA - Principal Component Analysis</i>	25
2. Objetivo	27
3. Material e Métodos	28
3.1 Material	28
3.2. Métodos	29
3.2.1 Construção do banco de dados	29
3.2.2 Desenho das estruturas	29
3.2.3 Cálculo dos descritores	29
3.2.4 Análise de dados	30
3.2.4.1 Modelos baseados em Regressão	30
3.2.4.2 Classificadores baseados em Árvore de Decisão	30
3.2.4.3 Modelos baseados em <i>Random Forest</i>	31
3.2.4.4 Modelos baseados em <i>MARS - Multivariate Adaptive Regression Splines</i>	32
3.2.4.5 Modelos baseados em <i>HCA e PCA</i>	33
4. Resultados e Discussão	34
4.1 Construção do banco de dados	34
4.2 Desenho das estruturas	36
4.3 Cálculo dos descritores	36
4.3 Curadoria dos descritores	37
4.4 <i>Forward e Backward</i>	38
4.5 Classificadores baseados em Árvore de Decisão	40
4.7 Modelos baseados em <i>Random Forest</i>	46
4.8 Modelo baseado em <i>MARS - Multivariate Adaptive Regression Splines</i>	53
4.9 <i>HCA e PCA</i>	58
5. Conclusões	66
REFERÊNCIAS	67
ANEXOS	76
Anexo A - Estruturas grupo fl e valores de <i>pKi</i> tanto para CB1R quanto CB2R	76
Anexo B - Estruturas e valores de <i>pKi</i> que atuam no CB1R	82
Anexo C - Estruturas e valores de <i>pKi</i> que atuam no CB2R	92

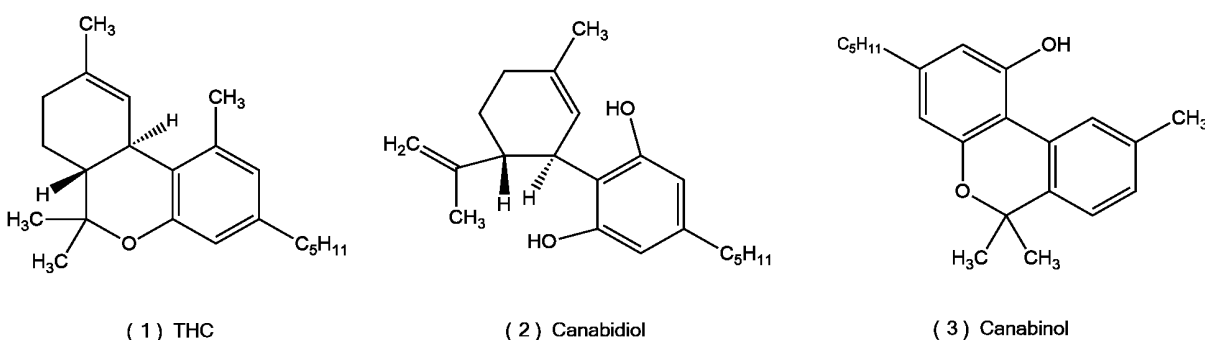
1. Introdução

1.1 Canabinoides e seus derivados sintéticos

Relatos sobre o uso terapêutico da *Cannabis sativa* L. remontam a diferentes períodos históricos e regiões do mundo, como o Egito, o Império Romano e a Índia. A primeira descrição farmacopeica dessa planta está relacionada à medicina tradicional chinesa, datando de cerca de 2000 a.C. Tais relatos atestam a longa e diversa utilização medicinal da *Cannabis sativa* L. para fins analgésicos e anestésicos (ADAMS; PEASE; CLARK, 1940).

As plantas do gênero *Cannabis*, tais como a *Cannabis sativa*, *Cannabis indica* e *Cannabis ruderalis*, possuem cerca de 400 compostos químicos distintos. Dentre esses compostos, destacam-se o canabidiol, o canabinol e o tetrahydrocannabinol (THC) como principais compostos bioativos. (Figura 01) (SILVA, 2009).

Figura 01 - Principais compostos bioativos do gênero *Cannabis*



Fonte: Elaborado pela autora

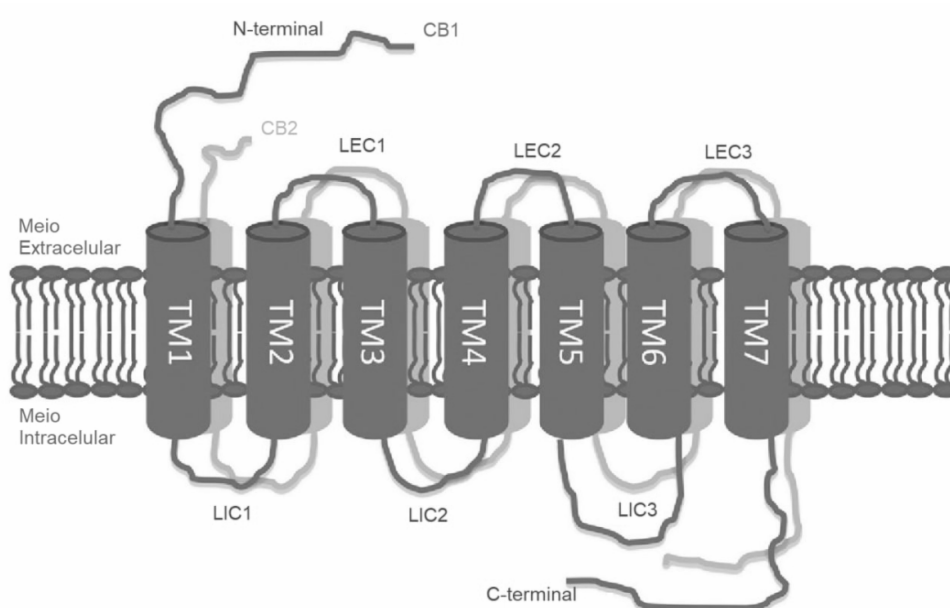
Na década de 1940, com o intuito de compreender melhor as substâncias bioativas responsáveis pelos efeitos psicoativos da planta, foram isolados o canabidiol e o canabinol. Entretanto, esses compostos não apresentaram essa atividade (ADAMS; PEASE; CLARK, 1940). Posteriormente, novos estudos indicaram que o responsável pelos efeitos psicoativos é o Δ9-tetrahydrocannabinol (THC), o qual foi isolado e descrito por Gaoni e Mechoulam em 1964 (PETERS *et al.*, 1999; SCHICHO; STORR, 2011).

Inicialmente, os efeitos relacionados à *Cannabis* eram considerados não-específicos. Contudo, foram descobertos dois receptores, os quais foram denominados de receptor canabinóide CB1R e CB2R pela NC-IUPHAR (*International Union of Basic and Clinical Pharmacology Committee on Receptor*

Nomenclature and Drug Classification) (MATSUDA *et al.*, 1990; MUNRO; THOMAS; ABU-SHAAR, 1993; SCHICHO; STORR, 2011). A partir dessas descobertas, o termo canabinoides passou a ser utilizado para descrever um grupo heterogêneo de substâncias, sejam elas naturais ou artificiais, que se ligam a esses receptores (SANTOS, 2018).

O receptor CB1 é predominantemente encontrado no sistema nervoso central (SNC), embora também seja encontrado em outros órgãos, como o fígado, tecido adiposo e pele (LU; MACKIE, 2021). Em contraste, o receptor CB2 é expresso principalmente em células do sistema imunológico (MUNRO; THOMAS; ABU-SHAAR, 1993; GALIÈGUE *et al.*, 1995), incluindo a microglia (STELLA, 2010; CABRAL; FERREIRA; JAMERSON, 2015), mas também podem ser expressos em neurônios (SPILLER *et al.*, 2019), tendo uma função anti-inflamatória (ZOU; KUMAR, 2018; TANAKA; SACKETT; ZHANG, 2020). Ambos receptores pertencem à família de receptores acoplados à proteína G (do inglês *G - Protein Coupled Receptors - GPCR*) de classe A (WEIS; KOBILKA, 2018), caracterizados por terem um N-terminal extracelular, sete alfa-hélices transmembranares (TM), três *loops* extracelulares (ECL), três *loops* intracelulares (ICL) e um C-terminal intracelular (LU; POTTER, 2017). Os receptores CB1R e CB2R têm uma sequência de aminoácidos com 44% de identidade e 68% de homologia entre as TMs (SHAHBAZI *et al.* 2020).

Figura 2 - Comparação da estrutura dos Receptores Canabinoides 1 e 2, em 2D.



Fonte: adaptado de LU; POTTER, 2017.

Os efeitos relacionados a estes receptores podem ser obtidos por agonistas sintéticos de receptores canabinoides conhecidos por "canabinoides sintéticos", que produzem efeitos similares aos da *cannabis* se ligarem aos receptores. Estes compostos foram desenvolvidos por pesquisadores que estudavam a ação da substância no corpo humano com o intuito de desenvolver fármacos para o tratamento de doenças neurodegenerativas, dependência de drogas, distúrbios da dor e do câncer, entre outras condições (EMCDDA, 2017).

Os compostos que possuem características canabinoides podem ser classificados com base nas estruturas químicas. Eles podem ser divididos em: (I) clássicos (THC, outros constituintes da *cannabis*; e seus derivados sintéticos estruturalmente relacionados); (II) análogos; (III) não clássicos; (IV) híbridos (combinações de características estruturais dos clássicos e não clássicos); (V) aminoalquilindóis; (VI) eicosanóides (endocanabinoides como a anandamida e seus análogos sintéticos), entre outros (UNODC, 2010).

Alguns canabinoides sintéticos foram avaliados em modelos celulares e ensaios *in vivo* com intuito de analisar o modo de ação e propriedades farmacológicas. Nestes estudos, encontrou-se grande dificuldade em dissociar as propriedades medicinais desejadas dos efeitos psicoativos indesejados, limitando, assim, as informações acerca de seus efeitos tóxicos no organismo. Apesar disso, é possível constatar que, quanto ao uso recreativo pode levar a manifestação de quadro psicoativo acompanhado de surtos de envenenamento, bem como outras alterações fisiológicas (SILVA, 2009).

A ação farmacológica dos canabinoides sintéticos pode ocorrer de maneira semelhante à *cannabis*, atuando como agonistas de receptores canabinoides ou antagonistas quando esses compostos se ligam aos receptores sem produzir os efeitos indesejados, porém bloqueando-os e impedindo a ligação de outros compostos (UNODC, 2010).

Encontrar moléculas que tenham ação seletiva nos receptores CB1R e CB2R representa um desafio científico (AN *et al.* 2020), uma vez que os sítios ortostéricos de ambos os receptores são significativamente semelhantes, podendo-se sobrepor à posição de agonistas (AN *et al.* 2020; HUA *et al.* 2020). Para auxiliar a busca por essas estruturas, podemos realizar a aplicação de técnicas computacionais que ao longo dos últimos anos, têm se tornado fundamentais para o desenvolvimento de

novos fármacos, especialmente em definições iniciais no *pipeline* do desenvolvimento de tais fármacos.

1.2 Quimioinformática

Quimioinformática é uma área que se utiliza de métodos de informática para resolver problemas originados na química, utilizando ferramentas como estatística, computação e matemática (GASTEIGER; FUNATSU, 2006). Sua aplicação facilita o planejamento e otimização de experimentos ao fornecer informações químicas relevantes para a análise dos dados obtidos, dando suporte à tomada de decisão no planejamento de fármacos. (FERREIRA *et al.*, 1999; CHEN *et al.*, 2018).

Estudos mostram que, à medida que as moléculas aumentam de tamanho o número de átomos e as possibilidades de ramificação aumentam, o que conduz à exponencial expansão do universo químico (BOHACEK; McMARTIN; GUIDA, 1996; SCHNEIDER, 2018). Como a busca exaustiva e aleatória se torna inviável, é lógico desenvolver estratégias racionais para avaliar o Espaço Químico para potenciais fármacos. Nesse contexto, há um movimento nas áreas de quimioinformática e química computacional voltado a estudar enormes bibliotecas químicas, empregando o uso sistemático de descritores químicos e representação visual do espaço químico a fim de explorar simultaneamente milhões de compostos (MEDINA *et al.*, 2022).

1.2.1 Descritores

Diversas ferramentas quimioinformatas foram desenvolvidas, como por exemplos, formatos de representação química, algoritmos de regressão e classificação, criação de bancos de dados químicos (quimiotecas) e descritores químicos. O objetivo deste último é transformar a estrutura molecular de uma substância e suas características em um conjunto de números gerados por meio de procedimentos lógicos e matemáticos, que são codificados a partir da representação da estrutura molecular da substância em questão (ALVES *et al.*, 2018)

Os descritores podem ser obtidos de maneira empírica e não empírica, esta corresponde aos índices topológicos e químico-quânticos (eletrônicos e geométricos, entre outros) e a primeira tem o objetivo de representar as propriedades físicas do composto (CONSONNI; TODESCHINI; PAVAN, 2002; ARRUDA, 2008).

Os índices ou descritores topológicos são os parâmetros numéricos que podem ser codificados e calculados a partir da conectividade e composição da

estrutura, ou seja, estrutura bidimensional. (PEREIRA, 2016). Já os índices ou descritores denominados eletrônicos podem ser obtidos a partir de cálculos mecânico-quânticos, os quais descrevem propriedades eletrônicas, características globais de distribuição de carga parcial, bem como a influência de certos grupos ou substituintes na densidade de distribuição eletrônica. Essas informações possibilitam a realização da interpretação físico-química do modelo matemático com o qual estão relacionados (KATRITZKY; GORDEEVA, 1993; PEREIRA, 2016).

Os descritores podem ser classificados de acordo com as suas dimensões estruturais ou função, como apresentado na Tabela 01.

Tabela 01 – Classificação geral do descritores estruturais

Representação Molecular	Função/ Exemplo	Exemplo
0D	Descritores Moleculares	Peso molecular Contagem dos átomos Ligações Número de carbonos Ligações rotacionáveis Entre outros
1D	Cômputo dos fragmentos Grupos funcionais	Contagem de hidroxilas, ácidos carboxílicos, carbonos Entre outros
2D	Índices de conectividade Índice de informações <i>2D matrix-based</i> <i>Auto-correlações 2D</i> Entre outros	<i>Information Content index</i> <i>ETA alpha</i> <i>Zagreb index</i> Entre outros
3D	Descritores geométricos <i>3D matrix-based</i> Autocorrelações 3D RDF 3D-MoRSE Entre outros	Excentricidade molecular Média do potencial eletrostático molecular Entre outros
4D*	3D	Média de diferentes confôrmeros
5D*	4D	Obtenção dos confôrmeros através de algum método de ajuste explícito (proteína)

Representação Molecular	Função/ Exemplo	Exemplo
6D*	5D	Obtenção dos conformeros através de diferentes graus de solvatação
Outros	Descritores de carga	Total da carga positiva e negativa Polaridade Entre outros
Outros	Propriedades moleculares	Índice de saturação Fator de hidrofiliidade Coeficiente de partição octanol-água Entre outros

* Nota: Alguns autores consideram a ausência de dimensão (0D) ou diferentes poses 3D como novas dimensões, não considerando as dimensões físicas.

Fonte: Adaptado de Oliveira (2015)

De maneira geral, o conhecimento de ferramentas quimiométricas e estatísticas é aplicado na análise e criação de modelos a partir de dados obtidos, sendo muito úteis na redução da dimensionalidade dos dados, o que facilita a interpretação do conjunto de informações (VALE, 2005).

1.2.2 Machine Learning

Machine Learning (do inglês, aprendizado de máquina) e a análise preditiva são técnicas que utilizam dados para realizar previsões, porém os processos utilizados são distintos. A análise preditiva tem como objetivo fazer estimativas sobre resultados futuros com base em dados históricos e técnicas de análise, utilizando ferramentas estatísticas para entender ocorrências futuras (RODRÍGUEZ-PÉREZ; BAJORATH, 2021). De outro lado, o aprendizado de máquina é um campo da inteligência artificial que se preocupa em desenvolver algoritmos e sistemas que possam aprender e melhorar com base em dados, sem serem explicitamente programados, baseando-se em padrões e inferências para fazer previsões e recalibrações conforme o conhecimento adquirido. Essa abordagem é capaz de detectar os padrões moleculares relacionados à atividade biológica e é amplamente utilizada tanto pela indústria quanto pela academia no contexto de descobrimento de fármacos (NEVES *et al.*, 2018).

No conjunto de ferramentas aplicadas a *Machine Learning*, desse conjunto de ferramentas, é possível citar diversos algoritmos de classificação e regressão que são corriqueiramente empregados, tais como: Regressão Linear Múltipla e Quadrados Mínimos Parciais (do inglês, *Partial Least Squares*), Análise Discriminante, k-Vizinho-Mais-Próximo (do inglês, *k-Nearest-Neighbor*), Máquinas de Vetores de Suporte (do inglês, *Support Vector Machines*), Árvores de Decisão (do inglês, *Decision Tree*), Florestas Randômicas (do inglês, *Random Forest*), Redes Neurais Artificiais (do inglês, *Artificial Neural Network*) entre outros (SVETNIK *et al.*, 2003).

1.2.2.1 Regressão Linear

Quando deseja-se um modelo ou função que relacione a resposta aos preditores, com o intuito de prever com precisão a resposta de observações futuras ou entender a relação entre a resposta e os preditores, utiliza-se um aprendizado chamado supervisionado. Este tipo de aprendizado é programado ou treinado a partir de um conjunto de dados pré-definidos ou rotulados. Como exemplo simples deste tipo de abordagem, temos a regressão linear, que tenta estabelecer a resposta quantitativa ao averiguar a relação linear entre uma ou mais variáveis: uma resposta (y) e uma preditora (X). Os coeficientes da equação, ou pesos, são calculados de acordo com a Equação (1).

$$w = (X^T X)^{-1} X^T y \quad \text{Equação (1)}$$

Caso o problema apresente um grande número de preditores (também chamadas de variáveis independentes), pode ocorrer *overfitting*. Para evitar esse problema, é necessário utilizar técnicas com intuito de selecionar os preditores mais relevantes, eliminando aqueles que não estão associados à(s) resposta(s). Para tal, como exemplo de metodologias empregadas, temos *Forward Selection* ou *Backward Elimination* (ENGEL; GASTEIGER, 2018).

A metodologia *Forward Selection* consiste em iniciar o modelo construindo uma equação que contenha apenas uma variável independente, avaliada como a que apresenta maior contribuição para predição, ou seja, maior ganho de informação. Em seguida, adiciona-se uma segunda variável a capacidade de predição do modelo e, de maneira subsequente, as demais variáveis são adicionadas, escolhendo-se aquelas que proporcionam maior ganho de informação

até que novas variáveis adicionais não interfiram significativamente o modelo (ENGEL; GASTEIGER, 2018). Por sua vez, a *Backward Elimination* realiza um processo inverso, em que se inicia o modelo incluindo todas as variáveis independentes, eliminando interativamente, uma a uma, até não se observar alteração significativa da qualidade do modelo (ENGEL; GASTEIGER, 2018). Ademais, pode ser realizado um modelo híbrido que, após adicionar cada nova variável, remove-se quaisquer variáveis que não forneçam mais uma melhoria no ajuste do modelo, tentando assim, realizar a melhor seleção do subconjunto, mantendo as vantagens computacionais da seleção de *Forward Selection* e *Backward Elimination*. (ENGEL; GASTEIGER, 2018)

Abordagens podem ser utilizadas para selecionar quais modelos apresentam maior ganho de informação, ou seja, se ajustam aos dados entre modelos com diferente números de variáveis, como por exemplo, *Mallows' Cp*, Critério de Informação Bayesiano (*BIC*, do inglês, *Bayesian Information Criterion*), Raiz Quadrada do Erro Médio (*RMSE*, do inglês, *Root Mean Square Error*), Soma Residual dos Quadrados (*RSS*, do inglês, *Residual Sum of Squares*) (JAMES *et al.*, 2011). O *BIC* avalia a probabilidade dos modelos posteriores e adiciona uma penalidade ao adicionar termos ao modelo, enquanto *Cp* de Mallows compara o modelo completo com os modelos com os subconjuntos de preditores, ajudando a obter um equilíbrio importante com o número de preditores no modelo (GILMOUR, 1996; NEATH; CAVANAUGH, 2011). Já *RSS* mede a discrepância entre os dados e um modelo de estimativa, e R^2 ajustado é a porcentagem de variação na resposta que é explicada pelo modelo, ajustada para o número de variáveis do modelo em relação ao número de observações (MONTGOMERY; PECK; VINING, 2012).

$$RSS = \sum_{i=1}^n (y_i - \hat{y})^2 \quad \text{Equação (2)}$$

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{Equação (3)}$$

$$R_{adj}^2 = \frac{(1 - R^2)(n - 1)}{n - p - 1} \quad \text{Equação (4)}$$

Onde n representa número de amostras analisadas, \hat{y} o valor predito pelo modelo, y é a média da variável resposta e p é o número de variáveis independentes (ou descritores neste caso) (MONTGOMERY; PECK; VINING, 2012).

A regressão linear é uma técnica estatística amplamente utilizada para modelar devido a sua facilidade de interpretação, permitindo sua aplicação em diversas situações, desde que os pressupostos básicos do modelo sejam atendidos. Além disso, ela permite analisar o impacto relativo de cada variável independente na variável resposta, bem como avaliar o ajuste do modelo e sua precisão na previsão de valores futuros. É uma técnica flexível que pode ser combinada com outras técnicas estatísticas, no entanto, pode ser sensível a *outliers* ou pontos extremos, que podem afetar significativamente os resultados do modelo. É importante considerar suas limitações e pressupostos antes de aplicá-la a um conjunto de dados específico.

1.2.2.2 Árvores de Decisão

A "árvore porfiriana", uma forma de árvore de decisão, é o tipo mais antigo conhecido de diagrama de árvore de classificação e foi concebida pelo filósofo grego Porfírio no século III d.C (DE VILLE, 2013). É considerado um dos primeiros algoritmos estatísticos a serem implementados em formato eletrônico durante a adoção de circuitos digitais para cálculos eletrônicos no século XX (DE VILLE, 2013; SONG; LU, 2015).

O algoritmo tem sido amplamente utilizado em diversas disciplinas devido à sua uma apresentação intuitiva em forma de árvore, o que melhora a compreensão e a disseminação dos resultados, podendo ser aplicadas a problemas de regressão e classificação (DE VILLE, 2013 SONG; LU, 2015). É reconhecido por sua facilidade de empregar, livres de ambiguidade e robustez que, mesmo na presença de valores ausentes, garante uma produção de resultados de alta qualidade que requer poucas suposições para implantação (DE VILLE, 2013; SONG; LU, 2015). Tem-se vários usos importantes, incluindo a seleção e a avaliação da importância relativa de variáveis; tratamento de valores ausentes; predição e manipulação de dados (SONG; LU, 2015). Todavia é importante considerar algumas limitações deste algoritmo, como o risco de sobre-ajuste (*overfitting*) e sub-ajuste (*underfitting*), particularmente quando se usa um conjunto de dados reduzido. Além do fato que uma forte correlação entre diferentes variáveis de entrada pode resultar na seleção

de variáveis que melhoram aparentemente as estatísticas do modelo, mas não estão relacionadas causalmente ao resultado de interesse (SONG; LU, 2015).

A Árvore de Decisão é uma técnica de modelagem que utiliza uma representação gráfica que se assemelha a um fluxograma, no qual escolhas e seus resultados são retratados em forma de uma árvore composta de nós, ramos e arestas (MAHESH, 2018). Cada nó representa um evento ou escolha, ou seja, um atributo em um grupo a ser classificado. Enquanto que os ramos representam os possíveis valores que o nó pode assumir e, as arestas do gráfico representam a decisão (MAHESH, 2018).

Do ponto de vista de validação, a Curva de Característica de Operação do Receptor (Curva *ROC - Receiver Operating Characteristic*) pode ser utilizada como critério para avaliação do modelo classificador (PROVOST; FAWCETT, 1997). Trata-se de uma técnica adequada para mensurar a capacidade preditiva de modelos de risco, baseando-se nos conceitos de sensibilidade e especificidade (DANTAS, 2013).. Esta análise é baseada na plotagem da taxa de verdadeiro positivo e da taxa de falso positivo, fornecendo uma pontuação para cada classificador, sendo um classificador ótimo aquele que mais se aproxima do número um (PROVOST; FAWCETT, 1997). A primeira indica que o elemento de entrada é Genuíno (Positivo) e o Classificador o classifica como Positivo, enquanto que a segunda, o elemento de entrada é Impostor (Negativo) e o Classificador o classifica como Positivo.(PROVOST; FAWCETT, 1997)

Árvore de Decisão, é amplamente utilizada em diversas áreas, incluindo medicina, biologia e engenharia, para avaliar a capacidade preditiva de modelos de classificação. É uma ferramenta confiável e eficiente para análise de risco financeiro, diagnóstico médico e tomada de decisões em negócios, desenvolvimento farmacêutico e finanças.

1.2.2.3 Random Forest

Um algoritmo bastante popular utilizado em problemas de classificação ou regressão é o *Random Forest*, baseado em árvores de decisão. Estas são classificadores atraentes devido à sua alta velocidade de execução, porém, possuem limitações na complexidade significando, geralmente, uma precisão abaixo do aconselhável nos dados de treinamento (HO, 1995). Assim sendo, Tin Kam Ho, em 1995, propôs pela primeira vez um método para construir classificadores

baseados em árvores em decisão cuja capacidade pode ser expandida arbitrariamente para aumentar a precisão de dados de treinamento e teste (HO, 1995). Com a abordagem proposta, se tornou possível aumentar a complexidade das árvores de decisão, melhorando a precisão dos resultados sem comprometer a velocidade de execução, possibilitando que o método se adapte de forma mais eficiente a diferentes cenários, oferecendo resultados mais precisos e confiáveis.

O princípio proposto foi a construção de múltiplas árvores em subespaços selecionados aleatoriamente do espaço de características como forma de melhorar a classificação combinada. Árvores em diferentes subespaços generalizaram sua classificação de forma complementar, podendo levar a uma melhora significativa na precisão dos resultados (HO, 1995). Logo em seguida, em 1997, Amit e Geman introduziram a ideia de pesquisar um subconjunto aleatório das decisões disponíveis ao dividir um nó, no contexto do crescimento de uma única árvore, contribuindo para o aprimoramento do método (AMIT; GEMAN, 1997). Posteriormente, outros pesquisadores aprimoraram o método implementando diferentes estratégias, como a seleção aleatória de nós para a realização de decisões, com intuito de propiciar melhores divisões ou gerar novos conjuntos de treinamento randomizando no conjunto de treinamento original. Esse tipo de técnica baseada num conjunto de modelos é conhecida como ensemble method (BREIMAN, 1999; DIETTERICH, 2000).

Random Forest é um algoritmo de aprendizado de máquina que utiliza o método de *bootstrap* das amostras para o treinamento. Isso significa que aleatoriamente é retirado um subconjunto de n amostras (com reposição) do grupo de treinamento, mantendo o tamanho original de dados. Isso significa que cada árvore de decisão é então treinada em um subconjunto de observações, em vez de utilizar todas as observações, tornando o modelo mais robusto e reduzindo o *overfitting*. Para cada grupo de amostras retirada, uma Árvore de Decisão é gerada, e em cada nó de decisão seleciona-se aleatoriamente um grupo de descritores definido pelo argumento "*mtry*". Este argumento traduz como o subconjunto de descritores que serão testados em cada nó de decisão, o que ajuda a reduzir a correlação entre as árvores e torna o modelo mais generalizável. (SVETNIK *et al.*, 2003).

Embora a *Random Forest* seja conhecida por sua rapidez em relação a outros modelos de aprendizado de máquina, a sua previsão pode ser mais lenta devido ao

processamento necessário de múltiplas árvores de decisão. Além disso, pode não funcionar bem em conjuntos de dados com muitas características redundantes, pois essas características podem prejudicar a eficácia do processo de seleção de recursos realizado pelo algoritmo (GENUER; POGGI; TULEAU-MALOT, 2010). No entanto, tem sido amplamente utilizado por sua facilidade de entendimento, uma vez que possui suas origens nas árvores de decisão, e pode ser aplicado tanto em problemas de classificação quanto de regressão. Além disso, é eficaz na estimativa de dados faltantes e mantém sua precisão mesmo quando uma grande parte dos dados estão faltando. Possui capacidade em lidar com a alta dimensionalidade dos dados, aliada à sua robustez a variáveis irrelevantes na resposta e também possibilitar a interpretação do modelo construído (BREIMAN, 2001).

1.2.3.4 MARS - *Multivariate Adaptive Regression Splines*

Em 1991, Jerome Friedman introduziu uma modelagem de regressão flexível de dados de alta dimensão conhecido como *MARS* (do inglês, *Multivariate Adaptive Regression Splines*) (FRIEDMAN, 1991). É uma metodologia que pode ser vista como uma generalização da regressão linear, podendo ser aplicada tanto para classificação quanto para regressão que envolvem um elevado número de entradas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Foi projetado para lidar com problemas de regressão não-linear multivariado, onde há mais de uma variável de entrada, onde a relação entre as variáveis de entrada e a de destino não é linear. Deste modo, o objetivo é descobrir um conjunto de funções lineares simples, cada uma com sua própria inclinação que caracterizam os dados, sendo possível combiná-las para fazer uma previsão (FRIEDMAN, 1991; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Enquanto a regressão ajusta em uma única linha reta os dados para modelar a relação entre as variáveis independentes e dependentes, *MARS* permite o uso de várias funções lineares, chamadas de *splines*, que se ajustam aos dados em diferentes subintervalos do espaço de entrada (FRIEDMAN, 1991). *MARS* também inclui transformações e interações dessas variáveis, permitindo uma melhor modelagem da relação entre elas. Além disso, o *MARS* emprega um processo de seleção de variáveis que identifica as mais importantes para a previsão e elimina as irrelevantes ou redundantes (FRIEDMAN, 1991).

MARS possui algumas limitações, incluindo a subjetividade na seleção de variáveis, o que pode afetar a interpretabilidade do modelo, especialmente quando há preditores altamente correlacionados. Além disso, pode haver o problema de *overfitting*, que pode resultar em uma pior generalização do modelo em dados não observados. Além de ser sensível a *outliers* e exigir uma curadoria cuidadosa dos dados antes da modelagem. (FRIEDMAN, 1991; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No entanto, os modelos MARS, são conhecidos por serem facilmente compreendidos e interpretados, permitindo lidar com dados contínuos e categóricos. Uma das principais vantagens deste algoritmo é a sua baixa exigência de preparação dos dados, podendo ser usado com conjuntos de dados numerosos (FRIEDMAN, 1991). Devido a sua flexibilidade, possuem alta capacidade de aprendizado, porém deve-se atentar com a seleção dos preditores, especialmente quando estão correlacionados, pois essa seleção arbitrária pode afetar a interpretabilidade dos resultados.

1.2.2.5 HCA - Hierarchical Cluster Analysis

A Análise Hierárquica de Agrupamentos (*HCA*) tem como propósito a análise da formação de agrupamentos em um conjunto de amostras, considerando a similaridade entre elas, ou com alguma outra propriedade de interesse. É um método de aprendizado não-supervisionado, o que significa que não requer uma variável resposta (SANTOS, 2009). Normalmente o resultado é representado em um gráfico bidimensional denominado dendrograma, no qual é possível observar as correlações e similaridades entre as amostras individuais (SILVA, 2009).

A similaridade entre as amostras é quantificada por meio de uma medida de distância, que pode ser calculada utilizando diferentes métodos, como centróide, ligação simples, ligação completa, e método de *Ward*. Neste trabalho optou-se pelo método de *Ward*, uma vez que minimiza a variação intra-grupos e maximiza a variação entre grupos, além de apresentar resultados sensíveis a *outliers*. Este método pode ser utilizado tanto para distâncias euclidianas quanto para outras distâncias (VALE, 2005).

Existem dois principais algoritmos hierárquicos que podem ser aplicados, sendo eles o aglomerativo (*AGNES*, do inglês, *Agglomerative Nesting*) e o divisivo (*DIANA*, do inglês *Divisive Analysis*), que operam de maneira inversa. *AGNES* é

ascendente, ou seja, inicialmente cada objeto é considerado como um *cluster* individual, em seguida, combina os dois *clusters* semelhantes em um novo *cluster* maior (denominados “nós”), de acordo com determinados critérios até que se atinja um platô (WANG *et al.*, 2018). Por outro lado, o *DIANA* segue uma abordagem oposta, partindo de cima para baixo, começando com a raiz, na qual todos os objetos são incluídos em um único *cluster*, em cada etapa da interação, o *cluster* mais heterogêneo é dividido em dois até que o número desejado seja alcançado (WANG *et al.*, 2018).

Para se determinar o número ideal de *clusters* de um conjunto de dados, existem diversos métodos. Alguns são baseados em critérios arbitrários, como o método do cotovelo (*Elbow*), enquanto outros são mais subjetivos, como o *Silhouette*. O primeiro, consiste em representar graficamente a variação dos dados em relação ao número de *clusters* e, escolher ponto de inflexão da curva (elbow) ou seja, o ponto em que o ganho em relação ao aumento do número de *clusters* diminui significativamente (JAMES; *et al.*, 2011; R Core Team, 2021) O método *Silhouette* mede quanto um ponto se encaixa em um *cluster*, e então é produzido um gráfico medindo a distância entre pontos de um *cluster* em relação aos pontos do outro mais próximo. O número ideal de clusters é encontrado quando a medida da Silhueta atinge seu valor máximo (ET-TALEBY; BOUSSETTA; BENSLIMANE, 2020; R Core Team, 2021).

É necessário ficar atento, pois a escolha do método de ligação e da métrica de distância pode afetar os resultados da *HCA* em quimioinformática (DUDEK; ARODZ; GÁLVEZ, 2006), além disso, a interpretação dos dendrogramas pode ser subjetiva (ERIKSSON *et al.*, 2008). Embora não forneça informações sobre a natureza das diferenças entre os grupos ou sobre os fatores responsáveis por essas diferenças, é útil por ser uma técnica não supervisionada, o que significa que não requer conhecimento prévio sobre as classes de amostras, podendo levar a novas descobertas (DUDEK; ARODZ; GÁLVEZ, 2006). Também pode ser usada para reduzir a dimensionalidade dos dados, permitindo uma melhor visualização e interpretação dos resultados e permite a identificação de *outliers*, que podem ser amostras incomuns ou erros experimentais (ERIKSSON *et al.*, 2008; DUDEK; ARODZ; GÁLVEZ, 2006).

1.2.2.6 PCA - *Principal Component Analysis*

PCA, sigla em inglês para Análise de Componentes Principais, é um método advindo da estatística usado extensivamente em quimiometria e quimioinformática para reduzir a dimensionalidade dos dados, sem perda de informações relevantes. Este método consiste na geração de combinações lineares de descritores originais, transformando variáveis correlacionadas em um número menor de variáveis não correlacionadas, chamadas de Componentes Principais (PC). Assim, maximiza-se a descrição da variância de dados ao condensar informações e facilitando a interpretação dos dados, usualmente em apenas duas ou três dimensões (WU, 1997; ENGEL; GASTEIGER, 2018). A aplicação de PCA é vantajosa em análises exploratórias de dados e é uma ferramenta essencial em muitas aplicações, tais como a análise de dados, descoberta de fármacos e estudos de QSAR.

Existem diversas abordagens para redução de matrizes de dados em componentes principais, dentre as disponíveis, destacam-se análise fatorial, análise fatorial de correspondência e mapeamento não linear (ENGEL; GASTEIGER, 2018). A identificação de padrões de similaridade nos dados, comumente referidos como *clusters*, é realizada por meio de gráficos e serve como base para a aplicação de outros métodos multivariados, como a modelagem independente para analogia de classes (SIMCA, do inglês, *Soft Independent Modeling by Class Analogy*) e a calibração, bem como a regressão em componentes principais (PCR, do inglês, *Principal Component Regression*). A PCR utiliza os componentes principais para a construção de um modelo linear (LOPES; LYRA, 2010; WU, 1997).

Como citado anteriormente, métodos classificatórios são amplamente empregados para correlacionar propriedades e características de conjunto de dados (compostos) com o objetivo de compreender e otimizar a atividade biológica que eles apresentam (CAVALCANTI, 2019). Desta forma, é possível construir representações matemáticas que discriminam atividade tóxicas, alucinógenas e, também, regulatórias do sistema imunológico apresentados para canabinoides naturais e sintéticos.

Essa é uma técnica útil para reduzir o número de variáveis em um conjunto de dados, identificando padrões e facilitando a análise e interpretação dos dados (JOLLIFFE, 2002). A normalização de dados reduz a influência de variáveis com escalas diferentes, tornando o PCA computacionalmente eficiente e adequado para grandes conjuntos de dados (ABDI; WILLIAMS, 2010). No entanto, sua aplicação

inadequada pode resultar em perda de informação e dificuldades na interpretação dos resultados. É importante escolher o número correto de componentes principais e estar ciente de que eles podem não ter interpretação direta ou intuitiva (JOLLIFFE, 2002).

2. Objetivo

Este estudo teve como objetivo a aplicação de técnicas de quimiinformática na construção de modelos capazes de reconhecer perfis estruturais de canabinoides sintéticos e diferenciá-los. Para isso, os seguintes objetivos específicos serão realizados:

- Construção de um banco de dados de completo e representativo de canabinoides sintéticos;
- Cálculo de descritores moleculares com as informações estruturais relevantes;
- Construção de modelos de quimiometria utilizando técnicas:
 - Árvores de Decisão
 - *HCA* (Análise Hierárquica de Agrupamentos)
 - *MARS* (Regressão de *Splines* Aditivas Multivariadas).
 - *PCA* (Análise de Componentes Principais)
 - *Random Forest*

3. Material e Métodos

3.1 Material

- Computador PC Windows, processador i7, 8 Gb de memória RAM, HD 1Tb, placa de vídeo ATI Radeon 2Gb.
- Servidor de ultra processamento SX 24U, 50GB RAM, 4 Tb HD, GPU Nvidia quadra.
- Máquina Virtual do VMWare ESXi 6.0 com Windows 10 instalado, 8 CPUs virtuais, 10GB de RAM, 1TB de HD.
- Servidor de processamento Dell PowerEdge T630, processadores 2x Intel Xeon E5-2650v3 (20 núcleos, 40 threads, 2.3GHz), 64GB de memória RAM, 12 TB de armazenamento em RAID 5 (8TB disponíveis para o usuário).
- *Softwares* utilizados:
 - MarvinSketch, versão 20.16.0, 2020 C
 - PaDel (Yap C.W. 2011)
 - Plataforma Knime 4.9.2
 - Python library pandas (v1.4.2)
 - RStudio *Software* Team (Versão 2022.12.0+353). Com as bibliotecas
 - caret (v 6.0-93)
 - cluster (v 2.1.2)
 - dendextend (v 1.14.0)
 - dplyr (v 1.0.7)
 - factoextra (v 1.0.7)
 - partykit (v 1.2-13)
 - parsnip (v 0.1.7)
 - plotly (v 4.10.0)
 - pROC (v 1.17.0.1)
 - purrr (v 0.3.4)
 - recipes (v 0.1.17)
 - rpart (v 4.1-15)
 - rpart.plot (v 3.0.9)
 - rsample (v 0.1.1)
 - stats (v 4.1.0)
 - tidymodels (v 0.1.4)
 - tidyverse (v 1.3.1)
 - tree (v 1.0-41)
 - workflows (v 0.2.3)

3.2. Métodos

3.2.1 Construção do banco de dados

Foi construído um banco de dados tendo como referência o artigo "*Discovery of High-Affinity Cannabinoid Receptors Ligands through a 3D-QSAR Ushered by Scaffold-Hopping Analysis*", de Giuseppe Floresta e colaboradores, publicado em 2018 no periódico *Molecules*. A publicação apresenta atividade biológica experimental de 327 estruturas de compostos canabinoides sintéticos de diferentes classes químicas tais como análogos de anandamida, benzoil / álcool-indóis, ciclohexil fenóis, dibenzofuranos, derivados de indazol, indazol-carboxilatos, imidazol-carboxamidas, indol-carboxilatos, indol-carboxamidas, naftoil-benzimidazoles, naftoil-indazoles, naftoil-indoles, naftoil-naftalenos, naftoil-pirroles e fenilacetil-indol (FLORESTA *et al.*, 2018).

Todos os compostos no banco de dados foram submetidos a um processo de curadoria manual. Nesse processo, foram removidos compostos com as seguintes características:

- Sais
- Misturas
- Compostos tautoméricos
- Compostos sem valores para atividade biológica
- Estruturas duplicadas que possuíam valores da atividade biológica (pKi) divergentes

3.2.2 Desenho das estruturas

A partir do banco de dados construído, desenhou-se as estruturas utilizando o *software* MarvinSketch®, sendo salvas em formato ".mol" a fim de garantir compatibilidade com ferramentas que calculam descritores físico-químicos (MARVIN, 2020).

3.2.3 Cálculo dos descritores

Após o desenho das estruturas que compõem o banco de dados, foram calculados os descritores moleculares no *software* gratuito *Padel-Descriptor* que calcula um total 1875 descritores (1D, 2D e 3D) e *fingerprints*. Dentre os descritores calculados, destacam-se o número de anéis aromáticos, número de doadores e aceptores de hidrogênio, além de descritores topológicos, eletrônicos, estéricos e quânticos (YAP, 2010).

3.2.4 Análise de dados

Optou-se por utilizar a linguagem de programação R (versão 2022.12.0+353) para conduzir as análises, construir os modelos (utilizando o pacote "*tidyverse*") e gerar gráficos (utilizando o pacote "*ggplot2*"), com o objetivo de comparar a performance preditiva e de classificação obtidas utilizando diferentes técnicas.

Realizou-se uma curadoria nos descritores, excluindo os que apresentavam valores constantes, variância com valor próximo a 0.0 e os altamente correlacionados entre si (multicolinearidade), devido à ausência de correlações significativas com os valores de pKi .

Após a seleção dos descritores, realizou-se uma normalização (Equação 5) dos dados por meio de uma padronização dos valores numéricos, a fim de se obter uma escala comum para construção dos modelos.

$$X = \frac{X - X_{\text{mínimo}}}{X_{\text{máximo}} - X_{\text{mínimo}}} \quad \text{Equação (5)}$$

3.2.4.1 Modelos baseados em Regressão

Devido ao grande número de preditores, realizou-se uma abordagem automatizada com objetivo de reduzir a matriz de dados, e para tal foram utilizadas as metodologias *Forward Selection* e *Backward Elimination*. Com a finalidade de verificar a acurácia e performance desses modelos, analisou-se as seguintes métricas: *RSS*, R^2 ajustado, *BIC* e *Cp de Mallows*.

3.2.4.2 Classificadores baseados em Árvore de Decisão

Realizada a curadoria inicial dos dados, foram selecionadas exclusivamente as estruturas que apresentavam dados experimentais de pKi para os ambos receptores canabinoídes. Com o objetivo de estabelecer uma prioridade de atuação do canabinóide, foi calculada a discrepância entre os valores experimentais de pKi dos receptores CB1R e CB2R. Com base nos resultados obtidos, estruturas com valores de diferença entre -0.1 e 0.1 foram consideradas imprecisas e, portanto, excluídas da análise.

Com os dados remanescentes, criou-se uma nova variável binária ou *dummy*, " pKi_{binario} ", que assume o valor de CB1R se a diferença for superior a 0 e CB2R caso contrário.

A partir disso, foi executado o modelo, gerando a árvore de decisão e sua representação gráfica, bem como informações adicionais, como o desvio médio residual, taxa de erro de classificação incorreta e as variáveis utilizadas na construção do modelo.

Com intuito de verificar a acurácia do modelo, dividiu-se os dados em dois grupos: teste (30%) e treino (70%). Em seguida, os valores de verdadeiro negativo e verdadeiro positivo foram somados para análise dos resultados obtidos.

Almejando aprimorar os resultados do modelo e efetuar as podas na árvore de decisão, realizou-se uma validação cruzada para determinar a complexidade preferível no modelo, uma vez que ela avalia a capacidade de generalização do modelo. Após a recomendação, o modelo foi executado novamente com os parâmetros sugeridos. Ao final, com propósito de comparar com a árvore original, foram verificados os mesmos parâmetros na árvore podada.

Estimou-se então a probabilidade de atuar em CB1R ou CB2R de cada nó da árvore. Por fim, realizou-se uma análise da Curva ROC (Característica de Operação do Receptor) para avaliação final do modelo.

3.2.4.3 Modelos baseados em *Random Forest*

Após a curadoria inicial dos dados, os preditores foram normalizados, centralizando e dimensionando os dados numéricos para obter um desvio padrão de um e média zero. Além disso, aplicou-se um filtro de alta correlação estabelecendo um limite para a correlação absoluta entre as variáveis e removendo aquelas com valores acima desse valor.

Realizada a curadoria dos dados, dividiu-se o *dataset* em dois subconjuntos: o conjunto treinamento com 70% e o conjunto teste com 30%. A estratificação foi realizada com base no índice de Tanimoto, utilizando amostragem aleatória para garantir a aleatoriedade da seleção dos dados. Este índice é uma medida utilizada para avaliar a similaridade entre dois vetores de características binárias, comumente empregado como medida de similaridade bidimensional entre duas estruturas. Um valor de índice de 0 indica nenhuma similaridade e 1 indica sobreposição completa dos dois conjuntos analisados. (TODESCHINI *et al.*, 2012)

Com intuito de garantir a reprodutibilidade dos resultados, é possível definir uma “*seeds*” utilizando a linguagem R, permitindo que as técnicas estocásticas sejam reproduzidas de forma consistente e comparável. Deste modo é possível

comparar a capacidade preditiva obtida utilizando técnicas diferentes, bem como manter a reprodutibilidade dos resultados.

Foi realizado um processo para selecionar "hiperparâmetros", que são parâmetros que controlam o próprio processo de treinamento do modelo (PROBST; WRIGHT; BOULESTEIX, 2019). Primeiramente foi criado um grupo randômico de amostras com suas "seeds" estabelecidas. Utilizou-se o algoritmo *ranger*, que ajusta um modelo criando um grande número de árvores de decisão independentes umas das outras. A previsão final é obtida por meio da combinação de todas as previsões das árvores individuais. Foram selecionados os hiperparâmetros que apresentaram menor erro após a validação cruzada. Em seguida, o modelo foi reajustado usando todas as observações disponíveis e o valor selecionado do parâmetro de ajuste. Com isso, obteve-se valores para o número de preditores para amostrar em cada divisão (*mtry*), número de observações necessárias para continuar dividindo os nós (*min_n*) e, o número de árvores (*trees*).

Realizou-se validação cruzada no grupo treino, criando assim, 10 subconjuntos dos dados. Nesse processo, foi reservado um subconjunto para realizar a validação, enquanto o modelo treinado com outros 9 subconjuntos. Este processo foi repetido até que todas as possibilidades fossem treinadas. Durante os testes, são calculados os componentes estatísticos e os resultados são pontuados. Após realizado o treino, foi realizada a validação dos resultados utilizando o grupo de estruturas separadas como "grupo teste".

Para avaliar a qualidade do modelo de regressão *Random Forest* em relação aos dados observados e sua capacidade de generalização para novos dados, utilizou-se duas métricas estatísticas para mensurar a qualidade: R^2 ajustado e *RMSE*. O primeiro, quanto mais próximo de 1 apresenta maior a capacidade do modelo em "explicar" os dados experimentais e, o segundo, que quanto menor o valor do *RMSE*, melhor é a capacidade preditiva do modelo.

3.2.4.4 Modelos baseados em MARS - *Multivariate Adaptive Regression Splines*

A construção de modelos baseados em *MARS* seguiu um procedimento semelhante ao utilizado para a obtenção dos modelos de *Random Forest*. Foram realizadas a normalização dos dados, a divisão randômica em grupo de teste e treino por meio de *seeds*, bem como a validação cruzada dos dados.

Durante a validação cruzada no grupo de treino, os resultados foram validados com o grupo de estruturas separadas como "grupo teste". Para avaliar os modelos de regressão construídos por *MARS*, foram utilizadas duas métricas estatísticas: R^2 ajustado e *RMSE*.

3.2.4.5 Modelos baseados em *HCA* e *PCA*

Após a curadoria inicial dos dados, aplicou-se um filtro para remover descritores que apresentavam correlação entre si acima de 0.55, estabelecido pelo coeficiente de correlação da metodologia de *Spearman*, que avalia a relação monotônica entre duas variáveis contínuas ou ordinais. Em seguida, os preditores foram normalizados, sendo centralizados na média e dimensionados para ter desvio padrão igual um.

Realizou-se uma análise de agrupamento utilizando o método *HCA* e análise da distância Euclidiana em que, maior o valor observado, menor a semelhança entre os descritores, e pelo método *Ward* (também denominado método da mínima variância). Foram utilizados tanto o algoritmo aglomerativo (*AGNES*) e o divisivo (*DIANA*), foram utilizados os métodos *Elbow* e *Silhouette* para realizar cálculos sugestivos para o número de *clusters*.

4. Resultados e Discussão

4.1 Construção do banco de dados

O banco de dados (*dataset*), utilizado para o estudo é proveniente do artigo intitulado: *Discovery of High-Affinity Cannabinoid Receptors Ligands through a 3D-QSAR Ushered by Scaffold- Hopping Analysis* de Floresta e colaboradores, publicado em 2018 (FLORESTA *et al.*, 2018). Nele foram reportados 327 compostos com seus respectivos valores de atividade biológica pKi (escala anti logarítmica da constante de dissociação), descrevendo a afinidade de ligação entre estrutura e os receptores. Destes, 312 compostos atuam no receptor canabinóide tipo 1 (CBR1) e 187 no receptor canabinóide tipo 2 (CBR2).

No processo de curadoria foram removidas estruturas químicas que apresentavam duplicidade, divergência nos valores de atividades, bem como misturas, sais, compostos tautoméricos e aqueles sem valores de pKi respeitando as boas práticas de curadoria de estruturas (BENTO *et al.*, 2020). Após a curadoria, restaram 172 substâncias que apresentam valores de atividade biológica simultaneamente para ambos receptores (grupo denominado doravante fl), 290 compostos que atuam no receptor CB1R (grupo CB1R), 175 em CB2R (grupo CB2R), conforme descrito na Tabela 02.

Tabela 02 – Número total de estruturas químicas antes e após realizar a curadoria para remover estruturas duplicadas, divergentes, misturas, sais e tautômeros.

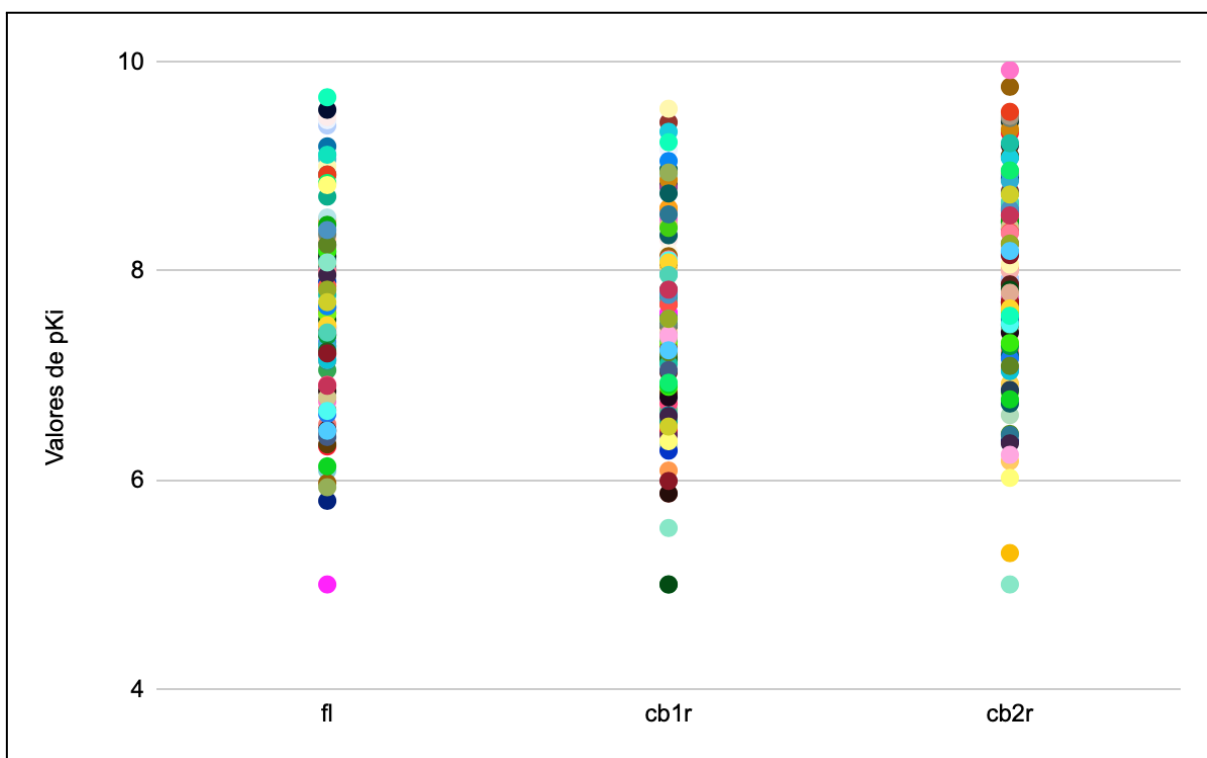
Grupo	Antes da curadoria	Pós curadoria	% de Exclusão
fl (CB1R e CB2R)	172	172	0.00 %
CB1R	312	290	7.05 %
CB2R	187	175	6.42 %
Total	327	294	10.1 %

Fonte: Elaborada pela autora.

Foi realizado um estudo de dispersão (Figura 03) para avaliar a distribuição dos compostos em relação a seus valores de pKi . Os resultados mostram uma homogeneidade na distribuição de valores, intervalo entre 5.00 e 9.92, ou seja,

houve uma distribuição maior que 3 unidades logarítmicas. Dessa forma, a variável resposta analisada está em conformidade com as boas práticas para o desenvolvimento de um modelo quantitativo de relação estrutura e atividade (TROPSHA, 2010).

Figura 03 - Gráfico de dispersão dos valores de pK_i dos compostos que atuam nos receptores CB1R e CB2R utilizados no estudo



Legenda: cb1_cb2 (fl) indica compostos com valores de pK_i para ambos os receptores canabinoides tipo 1 e 2. CB1R: compostos com valores de pK_i para o receptor canabinóide tipo 1. E, CB2R: compostos com valores de para o receptor canabinóide tipo 2.

Fonte: Elaborada pela autora.

O estudo destaca a importância da curadoria de dados em bancos de dados, que consiste na seleção criteriosa e remoção de estruturas químicas inconsistentes. Essa prática é fundamental para garantir a qualidade dos dados utilizados na análise, evitar distorções nos resultados e modelos desenvolvidos a partir dos mesmos. A homogeneidade na distribuição dos valores de pK_i dos compostos estudados, como evidenciado na análise da Figura 03, indica que a seleção foi criteriosa e bem distribuída, fornecendo uma base confiável para o desenvolvimento de modelos em quimioinformática.

O grupo CB1R apresenta um maior número de amostras em comparação com o grupo CB2R. Essa diferença pode ser atribuída à maior disponibilidade de estudos e interesse na pesquisa relacionada ao receptor CB1R, bem como à seletividade dos compostos em relação a cada receptor. Além disso, as características estruturais e funcionais dos receptores podem influenciar a disponibilidade de ligantes específicos e, conseqüentemente, a quantidade de dados disponíveis para cada receptor. Esses fatores contribuem para a disparidade no número de amostras entre os grupos CB1R e CB2R após a curadoria dos dados. No entanto, os resultados indicam a necessidade de uma abordagem equilibrada na obtenção de dados para ambos os receptores, a fim de promover uma análise mais completa e abrangente.

4.2 Desenho das estruturas

A partir do banco de dados (Anexo A, B e C), desenhou-se as estruturas bidimensionais utilizando o software MarvinSketch® e, posteriormente armazenadas em formato “.mol”, para possibilitar a utilização de ferramentas que realizam cálculos de descritores físico-químicos (MARVIN, 2020). O MarvinSketch® é uma ferramenta amplamente utilizada para desenho de estruturas moleculares e pode ser integrado a outros softwares para a realização de análises mais avançadas (ChemAxon, 2020).

A representação adequada das estruturas químicas é essencial para garantir a precisão dos cálculos de descritores físico-químicos e previsões de propriedades de compostos (TODESCHINI; CONSONNI, 2009). Erros de desenho podem afetar a determinação de descritores moleculares relevantes e, portanto, comprometer a qualidade dos modelos construídos. Por isso, é importante ter cuidado na representação das estruturas moleculares, a fim de minimizar erros que possam afetar a confiabilidade dos resultados obtidos.

4.3 Cálculo dos descritores

Após a representação adequada das estruturas dos compostos, o *software* Padel-Descriptor foi empregado para converter descritores moleculares numéricos a partir das estruturas químicas dos compostos em questão.

Foram calculados 1444 descritores moleculares 2D e 3D para cada um dos três grupos de estruturas (fl; CB1R; CB2R). Os descritores moleculares são parâmetros numéricos que descrevem as propriedades físico-químicas (variáveis

contínuas, como massa molecular e quantidade de átomos) ou estruturais (binárias, representadas por *fingerprints* que verificam a presença) (TODESCHINI; CONSONNI, 2009).

Diante dos descritores químicos calculados, é possível desenvolver modelos estatísticos capazes de estabelecer correlações entre as variáveis independentes (descritores químicos) e variável dependente (atividade biológica) (TODESCHINI; CONSONNI, 2009).

A seleção adequada de descritores moleculares é crucial para a precisão do modelo estatístico na correlação entre variáveis. Uma grande quantidade de descritores pode aumentar a complexidade e interpretabilidade do modelo, mas a seleção dos mais relevantes pode aprimorar a capacidade de previsão (TROPISHA; GRAMATICA; GOMBAR, 2003).

4.3 Curadoria dos descritores

Com intuito de selecionar as variáveis que resultem em um melhor modelo predição, reduziu-se o número de descritores a fim de obter uma melhor acurácia, minimizando o risco de não conseguir ter boa performance com grupos externos, aumentando a capacidade de generalização do modelo (GOODARZI, 2012). Portanto, foi restringido o número de descritores baseado na correlação linear entre si, a partir da matriz de correlação de *Spearman* (GOODARZI, 2012). Nessa matriz, são considerados os valores absolutos das correlações entre pares de descritores. Caso duas variáveis apresentem uma correlação elevada, a função analisa a correlação absoluta média de cada variável e elimina aquela que apresentar a maior média de correlação absoluta média (R Core Team, 2021).

Durante a análise da matriz de Spearman (Tabela 03), foi observado que alguns descritores apresentavam uma variância de correlação próxima ou igual a zero, assim, esses descritores foram removidos. Também foi constatada uma correlação elevada entre os descritores, pois mesmo ao tentar aplicar um corte com valores superiores a 0.80, ainda restaria um número expressivo de descritores (Tabela 03) o que tornaria essa abordagem pouco eficiente. Portanto, foi estabelecido o corte em 0.55 de correlação (r) entre os descritores.

Tabela 03 - Quantidade de descritores após a aplicação de critérios de exclusão

Números descritores restantes	CB1R	CB2R
Inicial	1444	1444
Exclusão de Variância > 0.0	960	929
Corte índice de correlação = 0.80	175	193
Corte índice de correlação = 0.55	62	50

Fonte: Elaborada pela autora.

Ao reduzir a alta correlação entre as variáveis preditoras, nesse caso os descritores, evita-se a multicolinearidade, que ocorre quando tais variáveis estão altamente correlacionadas entre si. Essa condição pode levar à insignificância matemática de um determinado descritor, mesmo que haja uma relação significativa entre ele e a atividade biológica. A presença de descritores altamente correlacionados pode afetar de forma significativa o modelo, levando a resultados equivocados em relação à resposta biológica (JAMES *et al.*, 2011; Minitab Support, 2019).

A alta correlação entre os descritores pode ser um desafio para a seleção de descritores eficientes para modelos preditivos, devido à multicolinearidade, que pode reduzir a eficácia do modelo e levar a resultados equivocados. A análise da matriz de *Spearman* é usada para avaliar a correlação entre descritores e reduzir a multicolinearidade. É essencial selecionar descritores com baixa correlação para garantir a eficiência e precisão do modelo. Outras técnicas, como a análise de componentes principais, podem ser usadas para lidar com a multicolinearidade e selecionar os descritores mais relevantes (JAMES; *et al.*, 2011).

4.4 Forward e Backward

Uma abordagem para selecionar descritores que estão associados às respostas é seguindo a metodologia *Forward Selection* e *Backward Elimination*. Para avaliar a acurácia e performance do modelo, considerou-se quatro parâmetros estatísticos: *RSS* (soma residual dos quadrados), R^2 ajustado, *BIC* (Critério Bayesiano de Schwarz) e *Cp* de Mallows.

Tabela 04 - Comparação com número mínimo de variáveis necessárias pelos métodos de decisão para análise de dados

	Método	RSS	R²adj	BIC	Cp
CB1R	<i>Forward Selection</i>	62	43	8	26
	<i>Backward Elimination</i>	62	44	17	31
CB2R	<i>Forward Selection</i>	50	42	12	20
	<i>Backward Elimination</i>	50	34	10	26

Fonte: Elaborada pela autora.

Observando as métricas obtidas na Tabela 04, constatou-se que a abordagem de *Forward Selection* é a mais adequada para os dados relacionados ao CB1R, fundamentado nos valores R² ajustado (43 variáveis) e *BIC* (8 variáveis). No primeiro critério, a indicação do uso de um pequeno número de descritores evidencia que a metodologia explicaria uma maior quantidade de informações. Já quanto ao segundo critério, o critério com menor número de descritores, indica uma maior adequação do modelo ao adicionar termos ao modelo (GILMOUR, 1996; NEATH; CAVANAUGH, 2011; MONTGOMERY; PECK; VINING, 2012).

Ao analisar as métricas para os dados relacionados ao CB2R apresentadas na Tabela 04, verifica-se que o método *Backward Elimination*, é o mais indicado, uma vez que este método possui valor reduzido de variáveis necessárias em relação ao R² ajustado (34 variáveis) e *BIC* (10 variáveis) e, o Cp de Mallows próximo do número de preditores (26 variáveis). Tais métricas sugerem que haveria um ajuste razoável do modelo aos dados e adequada explicação das informações, bem como possivelmente uma boa adequação ao adicionar novos termos ao modelo e com estimativas relativamente mais precisas e não-viciadas (GILMOUR, 1996; NEATH; CAVANAUGH, 2011; MONTGOMERY; PECK; VINING, 2012).

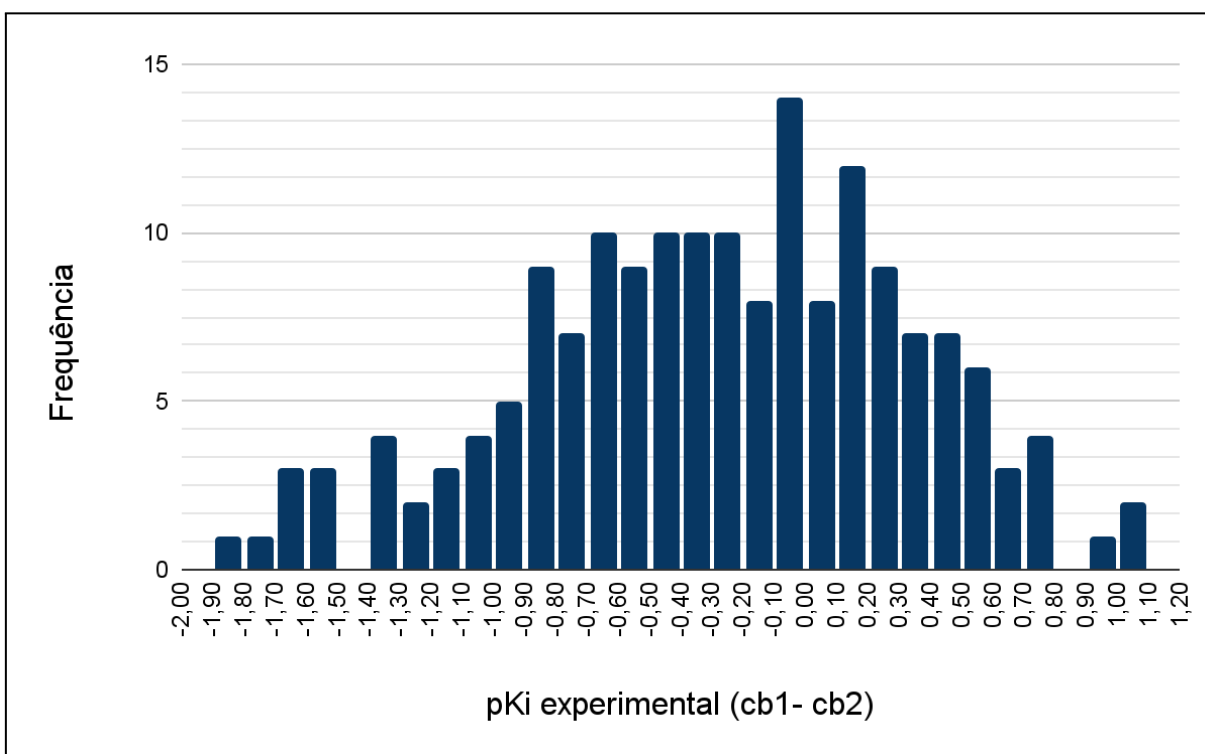
Empregar as metodologias a *Forward Selection* e *Backward Elimination* demonstrou ser uma abordagem eficaz para obter um modelo linear simples e facilmente interpretável. Essas metodologias fornecem uma maneira objetiva e reproduzível de reduzir o número de preditores em comparação com a escolha manual de variáveis, o que pode ser especialmente útil quando se busca um modelo com alta capacidade de interpretação e explicação.

A seleção apropriada de variáveis é um aspecto crítico na análise de modelos lineares, pois tem um impacto direto na acurácia, interpretabilidade e generalização do modelo. Um processo de seleção de variáveis adequado pode diminuir a complexidade do modelo, facilitando sua interpretação e aplicação prática. Contudo, a seleção do método apropriado depende das características dos dados e dos objetivos da análise, portanto, é fundamental selecionar o método mais apropriado para cada contexto específico.

4.5 Classificadores baseados em Árvore de Decisão

Foi realizada a curadoria inicial dos dados e calculado os descritores para as estruturas que possuíam dados experimentais de pKi para ambos receptores canabinoides, aqui chamado de grupo "fl". A variável dependente foi definida como a diferença entre os valores de pKi experimental do receptor CB1R e do CB2R. Os valores calculados variam de (- 1.90) a 1.08, como demonstrado na Figura 04.

Figura 04 - Histograma da distribuição de frequência dos valores resultantes da diferença entre os valores experimentais de pKi dos receptores CB1R e CB2R.

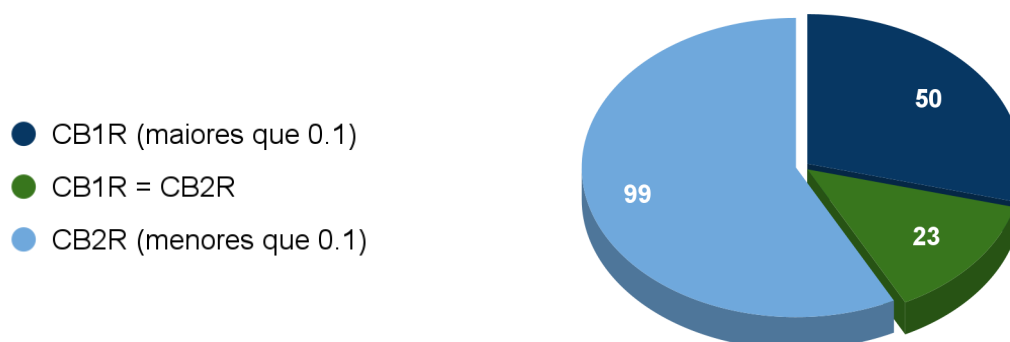


Fonte: Elaborada pela autora.

Deste modo, o processo de construção do modelo de árvore de decisão foi iniciado com a seguinte seleção de estruturas:

- 23 estruturas consideradas inespecíficas, cujos valores experimentais de pK_i para CB1R e CB2R foram considerados semelhantes, pois a diferença entre eles encontrou-se no intervalo de (-0.1) a 0.1;
- 50 estruturas com atuação preferencialmente em CB1R, ou seja, a diferença entre os valores experimentais de pK_i foi superior a 0.1;
- 99 estruturas com atuação preferencialmente em CB2R, ou seja, a diferença entre os valores experimentais de pK_i foi inferior a 0.1.

Figura 05 - Representação gráfica da distribuição proporcional das estruturas que atuam preferencialmente em cada receptor canabinoide.

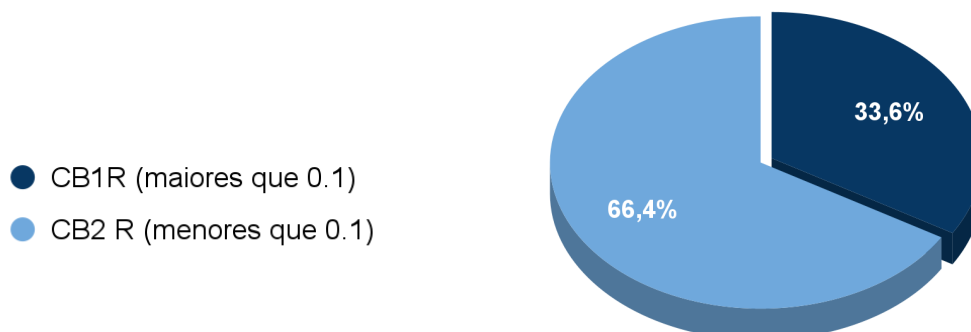


Fonte: Elaborada pela autora.

Os resultados experimentais com diferenças entre (-0.1) e 0.1 foram considerados como atuação imprecisa das estruturas em relação aos receptores canabinoides. Por esse motivo, para a construção do modelo de árvore de decisão, os seguintes compostos foram excluídos: fl_2, fl_11, fl_31, fl_38, fl_68, fl_73, fl_93, fl_94, fl_105, fl_121, fl_121, fl_136, fl_138, fl_143, fl_144, fl_146, fl_150, fl_153, fl_160, fl_166, fl_168 e fl_169 (Anexo A).

Após a exclusão destes valores, o modelo foi realizado com 149 estruturas, onde 34.8% dessas estruturas apresentaram preferência por CB1R, enquanto as demais, correspondentes a 65.2% apresentaram preferência por CB2R (Figura 06). A classificação foi realizada com base na diferença dos valores de pK_i experimental, onde valores superiores a 0 foram atribuídos ao CB1R e valores inferiores ou iguais a 0 foram atribuídos ao CB2R.

Figura 06 - Gráfico que ilustra a distribuição proporcional da afinidade preferencial por receptor canabinóide entre as estruturas que irão compor os modelos.

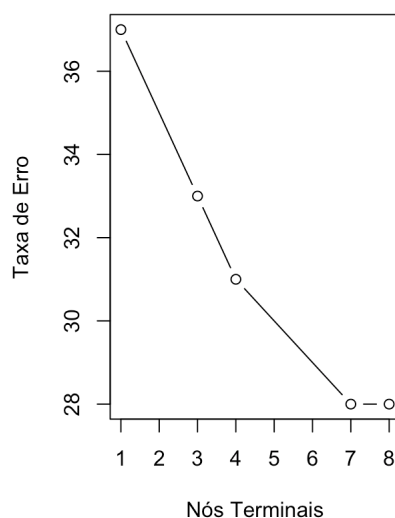


Fonte: Elaborada pela autora.

Dessa forma, iniciou-se o processo de construção de um classificador baseado em árvore de decisão com o objetivo de avaliar a viabilidade de distinguir as estruturas em relação aos receptores CB1R e CB2R a partir dos valores experimentais de pKi correspondentes.

O *dataset* foi dividido em dois grupos, 30% para teste e 70% para treinamento. Realizou-se validação cruzada com intuito de determinar a complexidade da árvore preferível, tendo como guia o erro de classificação. A Figura 07 representa essa análise, indicando que a menor Taxa de Erro foi alcançada com 7 Nós Terminais, totalizando 28 erros e uma taxa de classificação incorreta de 26.9%

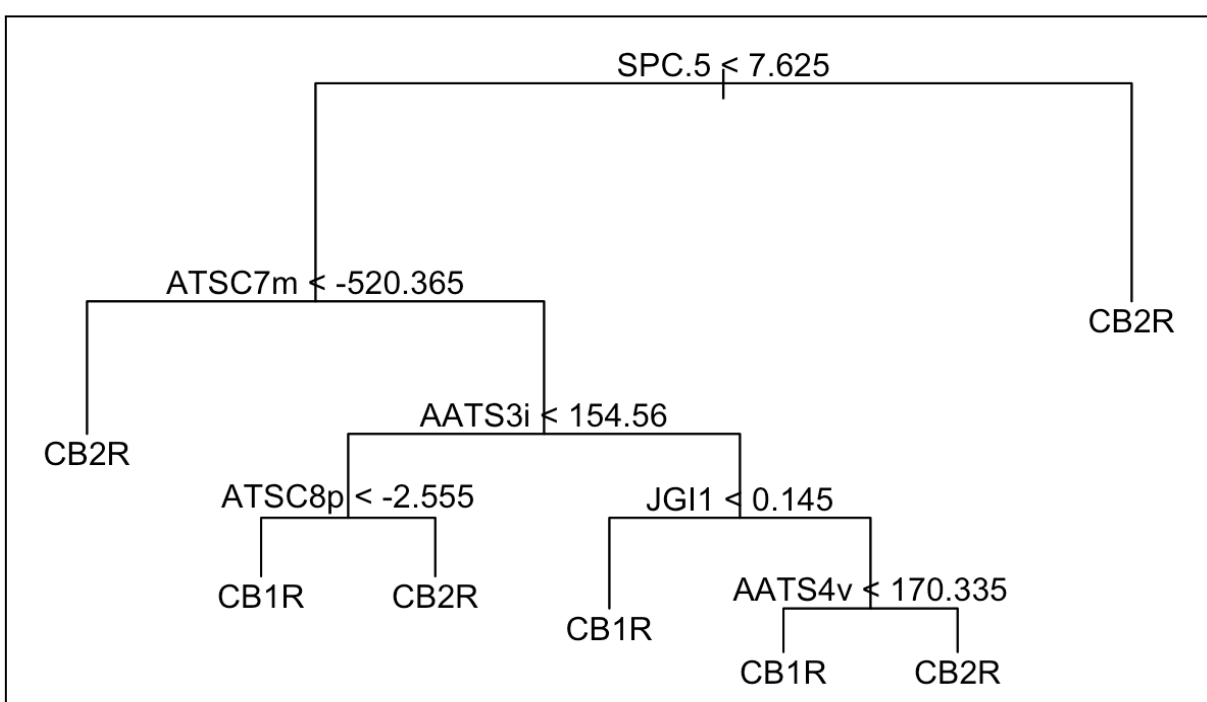
Figura 07: Ilustração do resultado da validação cruzada com intuito de determinar o número ideal de nós terminais na árvore de decisão, tendo como base a quantidade de erros de classificação.



Fonte: Elaborada pela autora.

Considerando a premissa que uma árvore ideal seria composta por 7 nós, selecionando os mais representativos que resultaram na árvore representada na Figura 08. Obteve-se um desvio médio residual com taxa em 0.2658 enquanto a taxa de erro de classificação incorreta = 0.0577, o que significa que o classificador previu incorretamente o resultado em apenas 5.8% das estruturas. Esse resultado indica que o modelo de classificação é capaz de prever os resultados da variável de forma robusta.

Figura 08: Árvore de Decisão gerada a partir do treinamento do modelo com dados experimentais de pK_i , utilizados para classificar entre CB1R e CB2R.



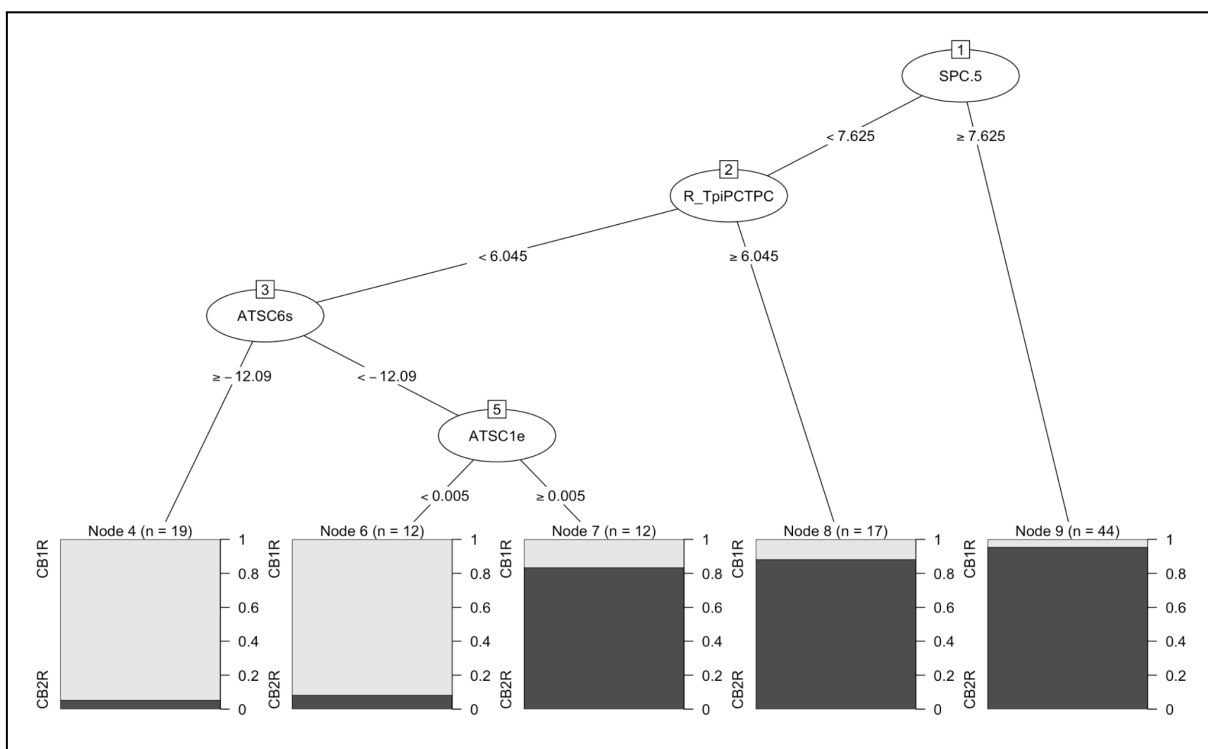
Fonte: Elaborada pela autora.

Na Figura 08 é possível observar os descritores ou variáveis utilizadas na construção do modelo: SPC.5, ATSC7m, AATS3i, ATSC8p, JGI1 e AATS4V. Nota-se que todos os descritores selecionados para realizar a classificação são 2D, sendo o principal deles (SPC-5) um descritor de índice Chi (*Chi path cluster - simple path cluster*) de conectividade molecular que codificam aspectos da estrutura molecular para uma molécula inteira (PaDEL-Descriptor, 2010).

Foi elaborada uma matriz de confusão para comparar os valores que um algoritmo alcançou como predição em relação aos valores reais ocorridos. Isso significa que somando o verdadeiro negativo e verdadeiro positivo, obteve-se 82. Esse resultado indica que, a cada cem classificações realizadas pelo modelo, ele foi

capaz de prever corretamente se uma estrutura atua em CB1 ou CB2 em 82% dos casos.

Figura 09: Representação gráfica da tendência de cada estrutura em atuar em um dos receptores CB1R e CB2R, com a indicação dos graus de probabilidade associados a cada nó terminal da árvore.



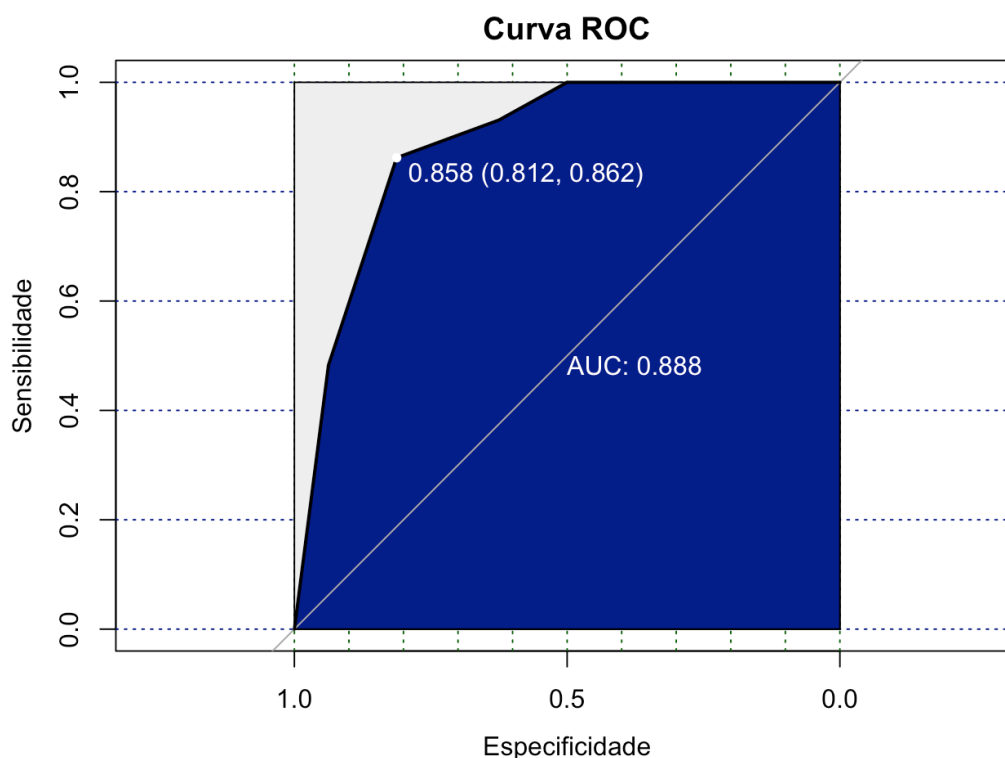
Fonte: Elaborada pela autora.

Na Figura 09 temos as probabilidades de uma estrutura atuar em CB1R e CB2R. Por exemplo, o nó #4 indica que cerca de 90% das escolhas do conjunto de treinamento que se enquadram nesse nó apresentam preferência por CB1R, quando a variável ATSC1e possui valores < 0.005 . Enquanto isso, no nó #9, as amostras que atendem aos critérios apresentam uma preferência de 90% para CB2R.

A curva ROC representa a relação entre a Taxa de Verdadeiro Positivo (TVP) e a Taxa de Falso Positivo (TFP) para diferentes limiares de classificação do modelo. É sabido que o modelo ideal deseja maximizar TVP (igual a 1) e minimizar o TFP (igual a 0), o que corresponde, visualmente, ao canto superior esquerdo da curva ROC. O modelo apresenta uma área sob a curva ROC (AUC - Area Under the Curve) de 0.888. Evidenciando um desempenho satisfatório do modelo de classificação binária na discriminação entre os receptores CB1R e CB2R, com base somente nos descritores moleculares utilizados. A alta sensibilidade e especificidade

obtidas pelo modelo reforçam sua capacidade de classificação precisa das amostras.

Figura 10 - Representação da curva ROC (*Receiver Operating Characteristic*) do modelo de classificação de Árvore de Decisão.



Fonte: Elaborada pela autora.

A curva ROC indica que o modelo que maximiza a Taxa de Verdadeiro Positivo ao mesmo tempo que minimiza a Taxa de Falso Positivo. Também podemos observar que a área sob a curva ROC (*AUC - Area Under the Curve*), é de 0.888. Desta forma, pode-se atestar que os dados de *pKi* experimental forneceram suporte para que o modelo de classificação binária fosse capaz de discriminar entre CB1R e CB2R apenas utilizando descritores moleculares.

Deste modo, estabeleceu-se o modelo de classificação baseado em árvore de decisão desenvolvido neste estudo, que mostrou-se eficaz na distinção entre as estruturas de canabinoides sintéticos que interagem com CB1R e CB2R, com uma taxa de erro de classificação incorreta de apenas 5.8%. A matriz de confusão e a curva ROC demonstraram a capacidade do modelo de prever com precisão as estruturas que atuam em cada receptor.

4.7 Modelos baseados em *Random Forest*

Pretendendo ampliar a gama de predições e considerando a presença de múltiplas variáveis no conjunto de dados, optou-se por utilizar o algoritmo *Random Forest*, o qual é reconhecido por sua capacidade em lidar com alta dimensionalidade (BREIMAN, 2001).

Realizada a curadoria e a normalização dos dados, dividiu-se o *dataset* em dois subconjuntos, 70% destinado ao conjunto treino (203 estruturas em CB1R, 122 em CB2R) e 30% para o grupo teste (87 estruturas em CB1R e 53 em CB2R). Essa divisão foi realizada por meio de amostragem randômica e com estratificação baseada nos valores de pKi . Segue abaixo os resultados obtidos para as estruturas que atuam em CB1R (Tabela 05) e em CB2R (Tabela 06).

Tabela 05 - Resultados das análises de *Random Forest* em estruturas que atuam no receptor CB1R: avaliação inicial, clusterização baseada no índice de Tanimoto, com ajuste de hiperparâmetros e, excluindo estruturas consideradas *outliers*.

Métrica	Grupo	Inicial	Cluster	Hiperparâmetros	Excluindo outliers
R²aj	Treino	0.948	0.938	0.944	0.956
	Teste	0.468	0.617	0.625	0.676
RMSE	Treino	0.272	0.290	0.274	0.268
	Teste	0.714	0.648	0.637	0.487

Fonte: Elaborada pela autora.

Tabela 06 - Resultados das análises de *Random Forest* em estruturas que atuam no receptor CB2R: avaliação inicial, clusterização baseada no índice de Tanimoto, com ajuste de hiperparâmetros e, excluindo estruturas consideradas *outliers*.

Métrica	Grupo	Inicial	Cluster	Hiperparâmetros	Excluindo outliers
R²aj	Treino	0.961	0.965	0.960	0.963
	Teste	0.593	0.641	0.648	0.678
RMSE	Treino	0.256	0.261	0.271	0.254
	Teste	0.714	0.626	0.618	0.549

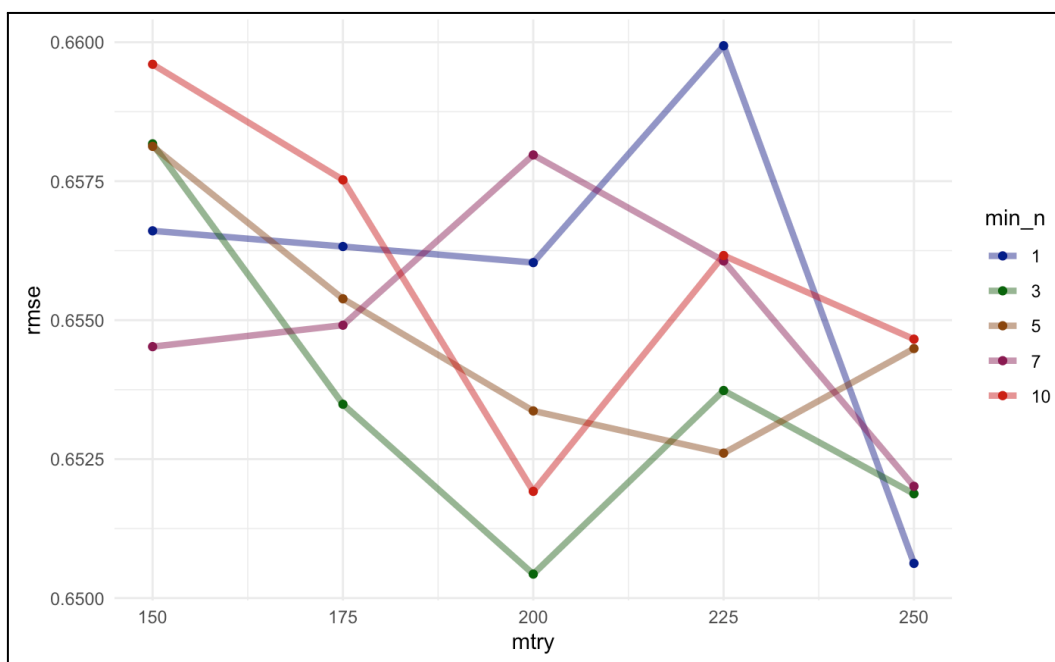
Fonte: Elaborada pela autora.

Ao analisar as Tabelas 05 e 06, constata-se a ocorrência de sobreajuste (*overfitting*) nos resultados apresentados na coluna "inicial", o que indica que os modelos apresentaram excelente desempenho no conjunto de treinamento, mas esse desempenho não se repetiu ao utilizar os dados do grupo teste.

Considerando esse achado, a primeira hipótese levantada foi a de que a divisão entre os grupos de teste e de treinamento não teria ocorrido de modo homogêneo. Com o intuito de verificar essa hipótese, procedeu-se com uma clusterização preliminar dos dados com base na estrutura, utilizando para isso o cálculo *fingerprints* e a distância de Tanimoto. Neste processo, as estruturas que atuam no receptor CB1R foram separadas em 29 grupos, enquanto as que atuam em CB2R foram divididas em 32 grupos. Em seguida, um novo modelo de *Random Forest* foi construído, agora com uma abordagem diferente na separação dos grupos de teste e treino em que a clusterização baseada na estrutura foi utilizada para separar os dados, com o objetivo de reduzir o risco de sobreajuste do modelo.

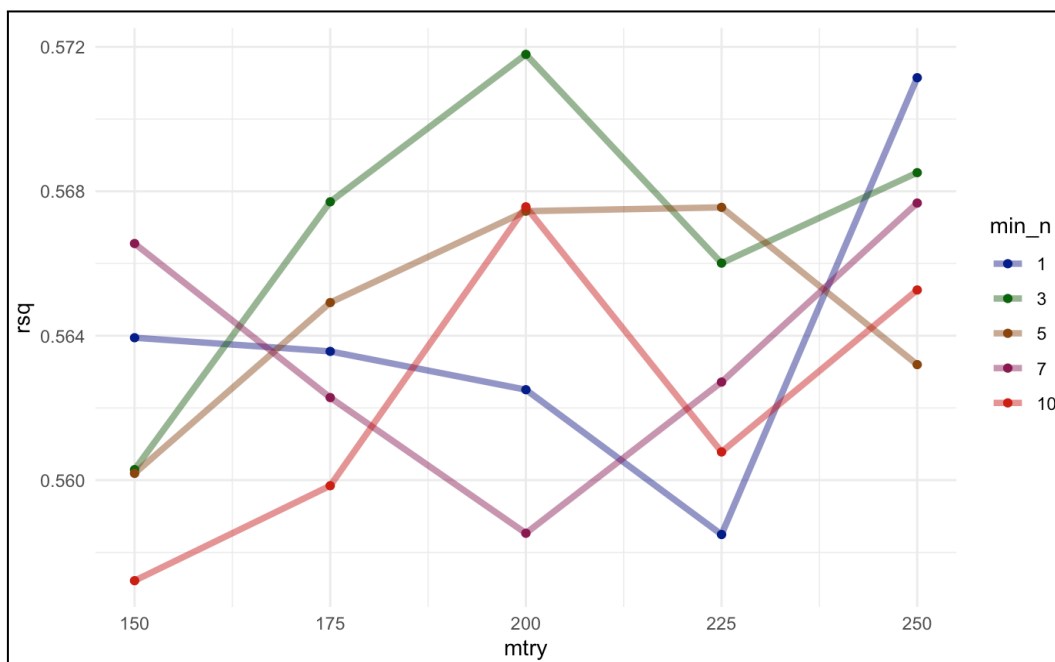
Com a finalidade de otimizar o desempenho do modelo, optou-se por selecionar os melhores valores para os "hiperparâmetros", chamados de parâmetros de ajuste, uma vez que seus valores devem ser otimizados com cuidado, por conduzirem à melhores resultados (PROBST; WRIGHT; BOULESTEIX, 2019). Foram selecionados os hiperparâmetros "*min_n*" e "*mtry*", que são utilizados para o ajuste fino em *Random Forest*. Sendo que "*min_n*" estabelece o número mínimo de observações em cada nó da árvore para evitar *overfitting*, enquanto "*mtry*" define o número de variáveis aleatórias selecionadas em cada divisão de nó, com o intuito de aumentar a eficiência do modelo e prevenir sobre-ajuste. Para isso, foi criado um grupo de amostras randomicamente por "*seeds*", para isso foi utilizado um *grid* de 15 pontos para selecionar os melhores valores, que foram posteriormente refinados para uma análise mais aprofundada. Os resultados obtidos foram representados nas Figuras 11 a 14.

Figura 11 - Gráfico de otimização do hiperparâmetro min_n (1-10) e $mtry$ (150-250) com base em dados $RMSE$ para o modelo de *Random Forest* aplicado ao CB1R



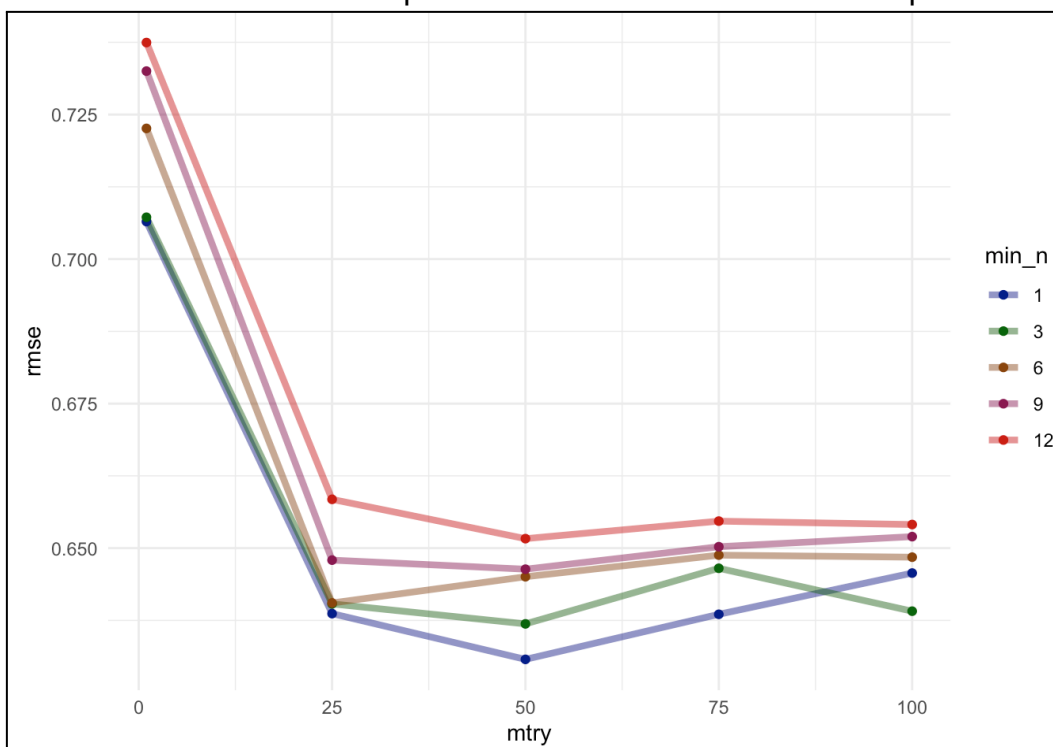
Fonte: Elaborada pela autora.

Figura 12 - Gráfico de otimização do hiperparâmetro min_n (1-10) e $mtry$ (150-250) com base em dados RSq para o modelo de *Random Forest* aplicado ao CB1R



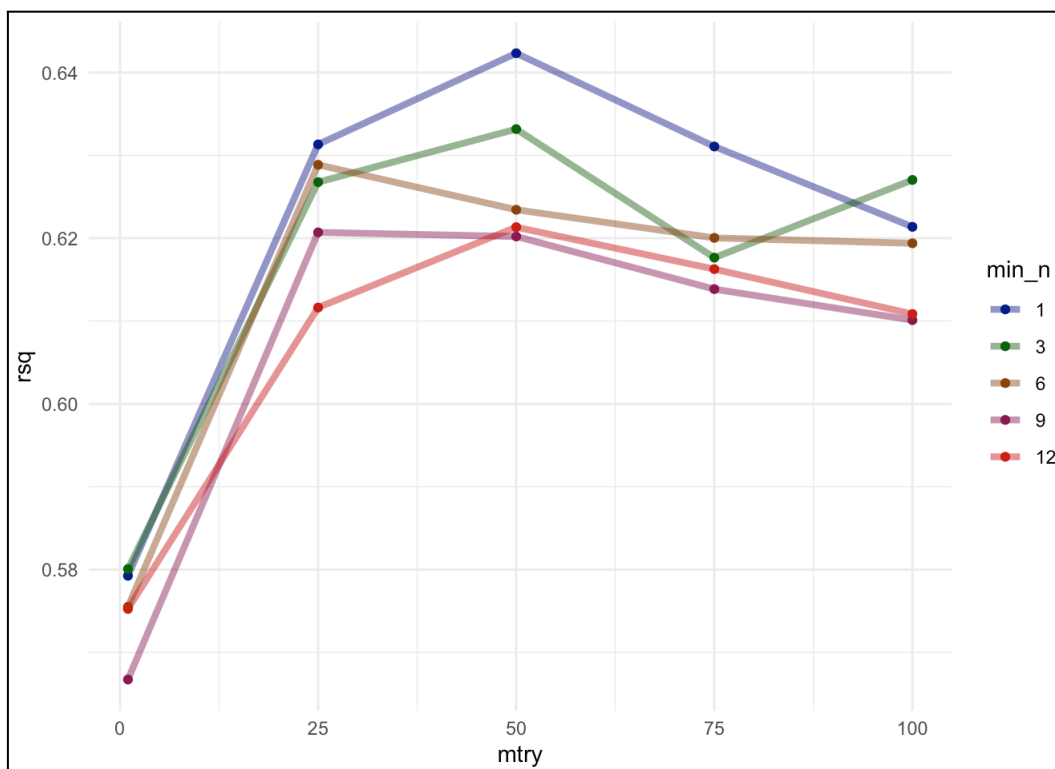
Fonte: Elaborada pela autora.

Figura 13 - Gráfico de otimização do hiperparâmetro min_n (1-10) e $mtry$ (1-100) com base em dados $RMSE$ para o modelo de *Random Forest* aplicado ao CB2R



Fonte: Elaborada pela autora

Figura 14 - Gráfico de otimização do hiperparâmetro min_n (1-10) e $mtry$ (1-100) com base em dados RSq para o modelo de *Random Forest* aplicado ao CB2R

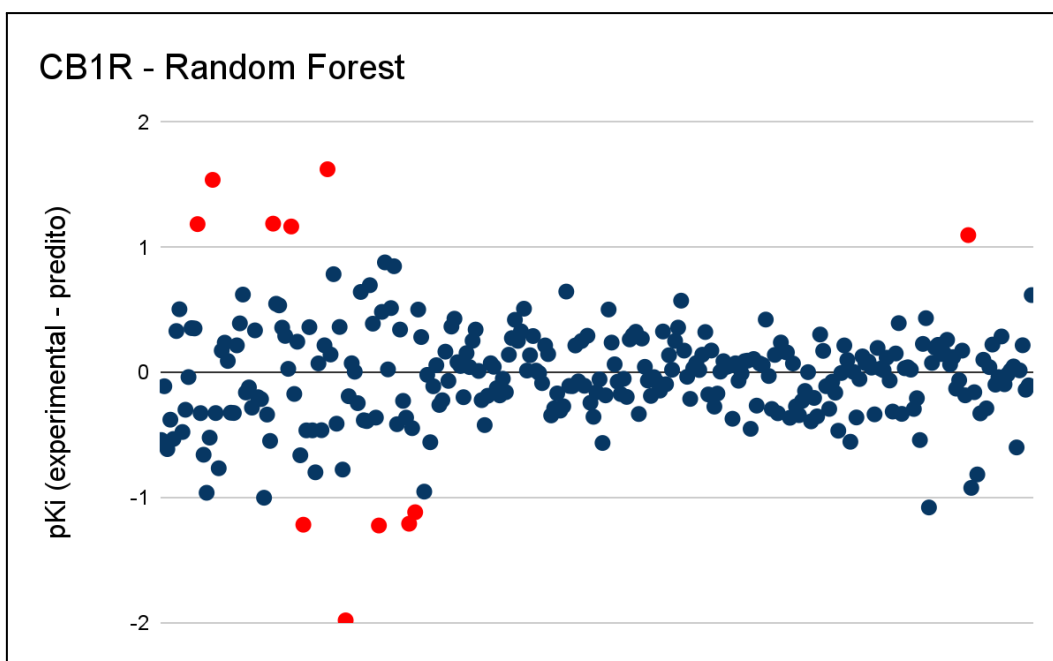


Fonte: Elaborada pela autora

Na Figura 11 e 12 é possível observar que a otimização dos hiperparâmetros $mtry$ e min_n para o modelo de *Random Forest* aplicado ao CB1R, com valores refinados entre 150-250 e 1-10, respectivamente, resultou no melhor desempenho quando $mtry = 150$, $min_n = 3$, e $trees = 1000$, resultando em melhores resultados para $RMSE$ e RSq . Já a análise da Figura 13 e 14 demonstra que os melhores valores comuns para $RMSE$ e RSq foram obtidos com $mtry = 25$, $min_n = 6$ e $trees = 1000$, ao selecionar valores de $mtry$ entre 1 e 100 e min_n entre 1 e 12 para CB2R.

A análise dos resultados da seleção mostrou uma diminuição do sobreajuste dos modelos apresentados na coluna "hiperparâmetros" da Tabela 05. No entanto, considerando possível aprimorar ainda mais o desempenho, optou-se por explorar uma outra estratégia por meio da exclusão de algumas estruturas que poderiam estar contribuindo para tal efeito. Para isso, foi realizado o gráfico de resíduo (a diferença entre os valores da diferença entre o pKi experimental e o predito) dos que atuam em CB1R. Verificamos que aproximadamente 4% (12 estruturas) apresentaram diferenças superiores a 1, conforme demonstrado na Figura 13.

Figura 15: Gráfico de resíduos com os valores da diferença entre pKi preditos por *Random Forest* e os valores experimentais para estruturas que atuam em CB1R.



Legenda: Pontos em vermelho foram consideradas estruturas *outliers*.

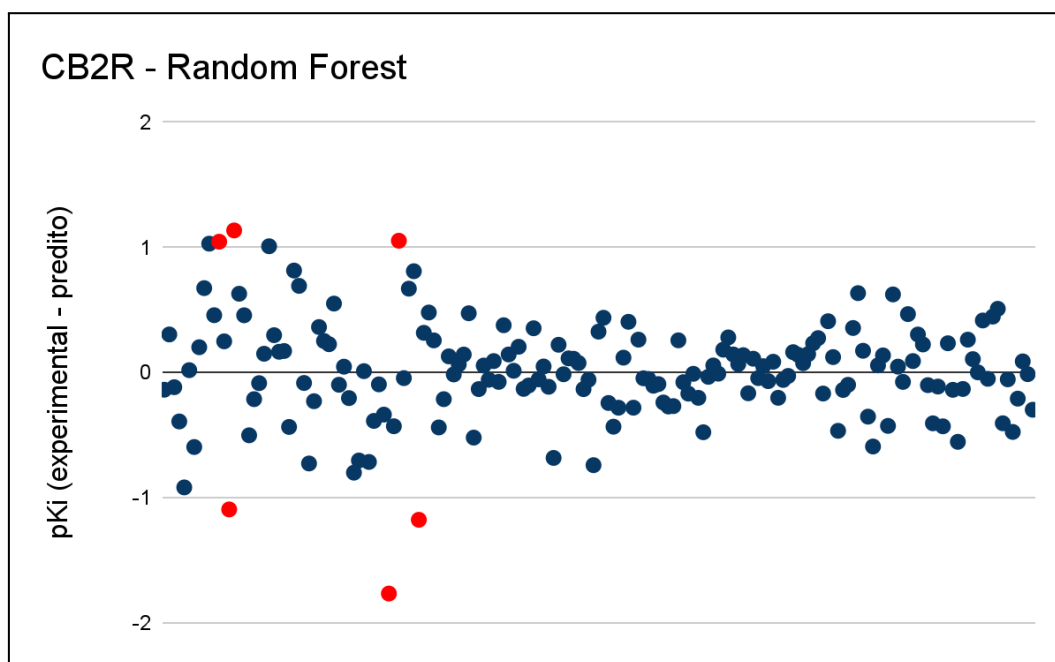
Fonte: Elaborada pela autora

Ao excluir estruturas identificadas como *outliers* (CB1R_42, CB1R_45, CB1R_59, CB1R_75, CB1R_99, CB1R_118, CB1R_137, CB1R_147, CB1R_173,

CB1R_188, CB1R_236), representadas em vermelho na Figura 13, verifica-se uma melhora significativa dos resultados e na redução do *overfitting* como atesta a coluna "excluindo *outliers*" da Tabela 05.

Da mesma forma, foi analisado o gráfico de resíduo dos que atuam no CB2R. Verificamos que aproximadamente 8 estruturas (2.75% do total) apresentavam uma diferença superior a 1.0 conforme podemos verificar na Figura 16:

Figura 16 - Gráfico de resíduos com os valores da diferença entre pKi preditos por *Random Forest* e os valores experimentais para estruturas que atuam em CB2R.



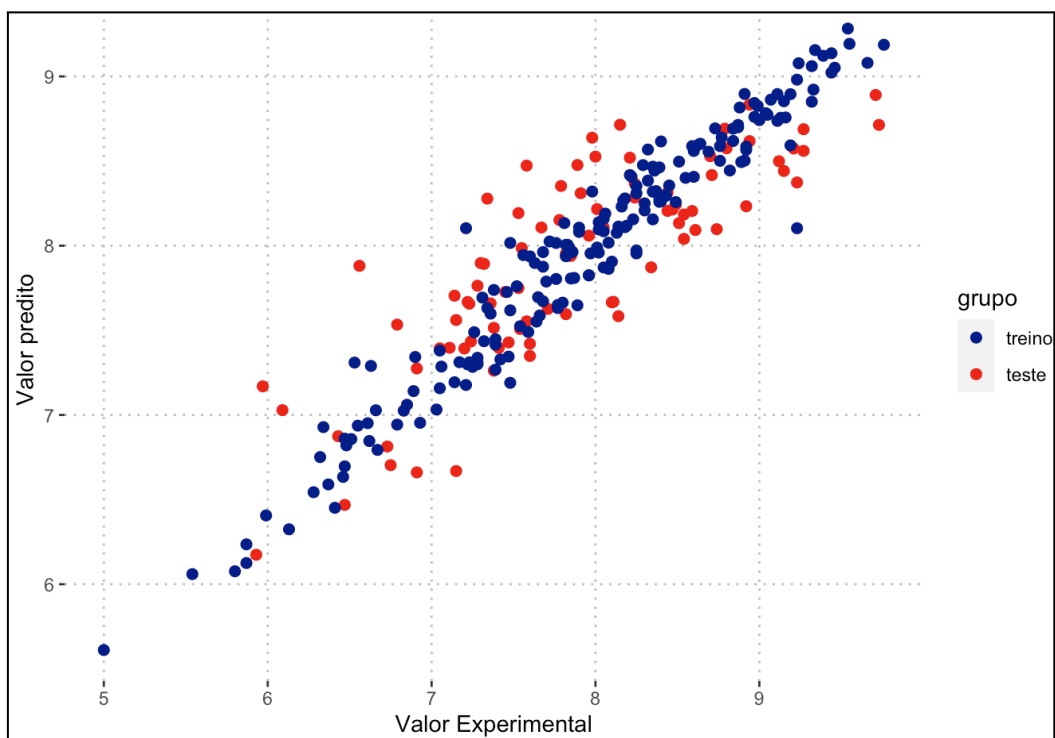
Legenda: Pontos em vermelho foram consideradas estruturas *outliers*.

Fonte: Elaborada pela autora

Analisando os dados, percebeu-se que retirando as estruturas CB2R_28, CB2R_33, CB2R_41, CB2R_149, CB2R_158, CB2R_165 (representadas em vermelho na Figura 16), resultou em uma melhora nos parâmetros avaliados e uma redução significativa de sobre-ajuste (Tabela 06 - coluna "excluindo *outliers*").

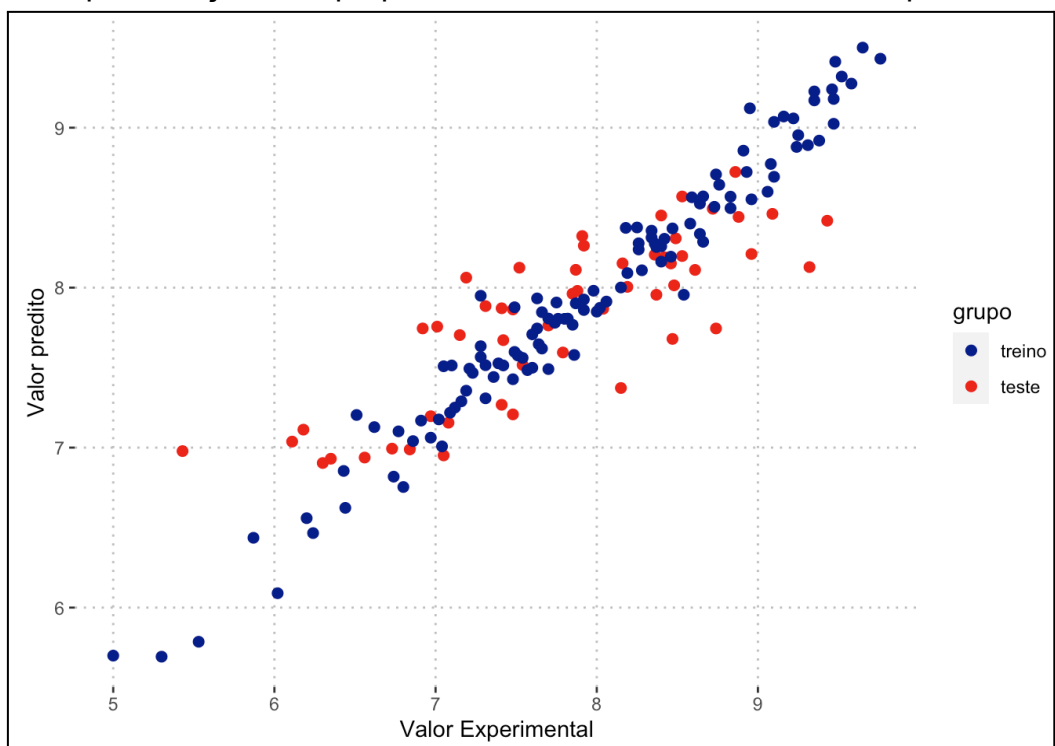
A otimização dos modelos, realizado por meio das mudanças realizadas, levou a melhorias consideráveis. Figuras 17 e 18 mostram as comparações entre os valores de treinamento e teste obtidos com o método *Random Forest* utilizando todos os descritores para o receptor CB1R e CB2R, respectivamente.

Figura 17 - Análise comparativa dos desempenhos do grupo teste e treino do modelo *Random Forest* aplicado em estruturas moleculares que interagem com CB1R, após seleção de hiperparâmetros e exclusão de valores atípicos.



Fonte: Elaborada pela autora.

Figura 18 - Análise comparativa dos desempenhos do grupo teste e treino do modelo *Random Forest* aplicado em estruturas moleculares que interagem com CB2R, após seleção de hiperparâmetros e exclusão de valores atípicos.



Fonte: Elaborada pela autora.

A abordagem do *Random Forest* mostrou-se computacionalmente eficiente, proporcionando resultados confiáveis e interpretação viável. Além disso, a metodologia demonstrou versatilidade na análise de dados moleculares e apresentou um bom desempenho na predição de resultados. Para melhorar a capacidade de generalização do modelo e evitar o *overfitting*, foram realizadas várias tentativas, como a otimização dos hiperparâmetros e clusterização dos dados. Essas estratégias ajudaram a reduzir o viés do modelo, aumentando sua capacidade de se adaptar a diferentes conjuntos de dados e melhorando a acurácia de previsão em dados não observados.

4.8 Modelo baseado em *MARS* - *Multivariate Adaptive Regression Splines*

Almejando empregar outra abordagem preditiva, porém de modelagem de regressão flexível de dados de alta dimensão conhecida, optou-se por aplicar a metodologia *MARS*.

Para realizar a construção de modelos utilizando *MARS*, foram realizadas etapas semelhantes às empregadas em *Random Forest*, incluindo a normalização dos dados, a divisão randômica dos grupos de teste (30%; composto por 87 estruturas em CB1R e 53 em CB2R) e de treinamento (70%; composto por 203 estruturas em CB1R e 122 em CB2R) por meio de determinação de *seeds* e a validação cruzada dos dados. A avaliação dos modelos de regressão construídos por *MARS* foi realizada utilizando-se métricas estatísticas R^2 ajustado e *RMSE*.

Os resultados obtidos para as estruturas que atuam em CB1R (Tabela 07) e CB2R (Tabela 08) estão apresentados a seguir:

Tabela 07 - Análise de dados obtidos por meio do método *MARS* para estruturas que atuam no receptor CB1R, incluindo análises inicial, com clusterização baseada no índice de Tanimoto e exclusão de *outliers*.

Métrica	Grupo	Inicial	Cluster	Excluindo outliers
R²aj	Treino	0.739	0.666	0.772
	Teste	0.486	0.512	0.502
RMSE	Treino	0.422	0.537	0.463
	Teste	0.771	0.721	0.657

Fonte: Elaborada pela autora.

Tabela 08 - Análise de dados obtidos por meio do método *MARS* para estruturas que atuam no receptor CB2R, incluindo análises inicial, com clusterização baseada no índice de Tanimoto e exclusão de *outliers*.

Métrica	Grupo	Inicial	Cluster	Excluindo outliers
R²aj	Treino	0.856	0.848	0.835
	Teste	0.554	0.567	0.362
RMSE	Treino	0.360	0.388	0.417
	Teste	0.763	0.659	0.867

Fonte: Elaborada pela autora.

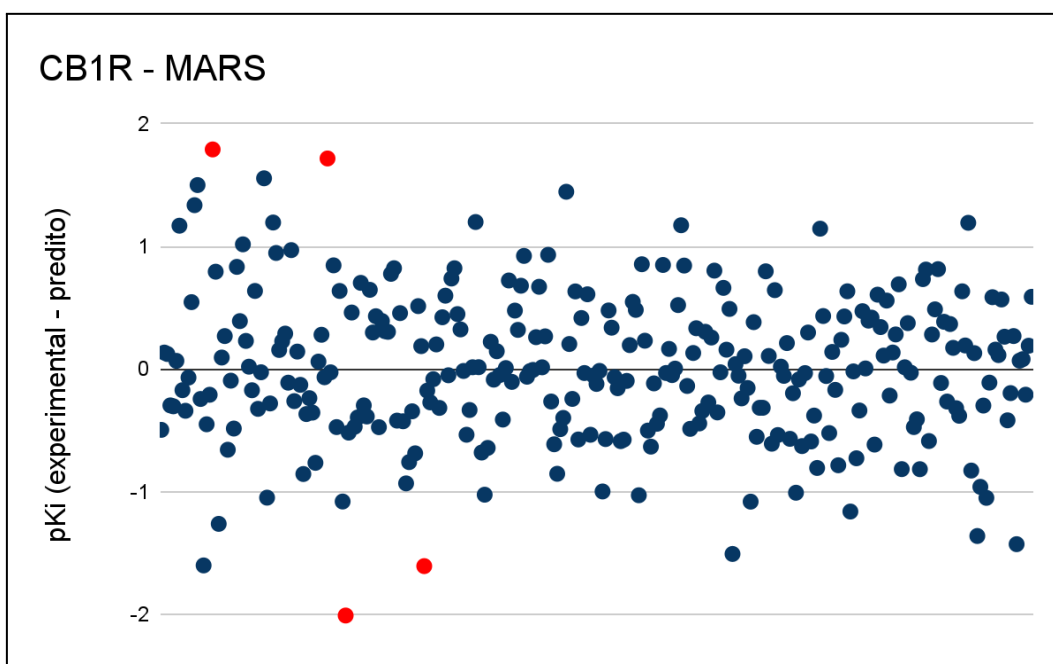
Explorando as tabelas 07 e 08, observa-se que na coluna "inicial" uma possível ocorrência de sobreajuste (*overfitting*) em ambos os modelos, uma vez que desempenho obtido no treinamento foi excelente, porém quando se utilizou os dados de grupo teste, esse desempenho não foi repetido.

Ocorrido isto, supôs-se que a divisão dos grupos testes e treinos não tenha ocorrido de forma homogênea. Para abordar essa questão, realizou-se uma prévia clusterização dos dados com base na estrutura, utilizando-a para o cálculo dos *fingerprints* e a distância de Tanimoto, conforme o realizado no modelo de *Random*

Forest. Com os dados já clusterizados, aplicou-se o modelo de *MARS* baseando a separação dos grupos testes e treino nos novos grupos, com objetivo de reduzir o sobre-ajuste. Ao analisar os dados encontrados (Tabela 07 e 08, coluna "*cluster*"), porém não percebeu-se uma melhora significativa nos resultados.

Visto isso, consideramos que alguns compostos poderiam não estar em consonância, e, portanto, decidimos excluí-los da análise. Para isso, analisamos a diferença entre o *pKi* experimental e o predito dos que atuam em CB1R. Verificamos que cerca de 14 estruturas (4.8%) apresentavam uma diferença superior a 1.0 conforme pode-se verificar na Figura 19:

Figura 19 - Gráfico de resíduos com os valores da diferença entre *pKi* preditos por *MARS* e os valores experimentais para estruturas que atuam em CB1R.



Legenda: Pontos em vermelho foram consideradas estruturas *outliers*.

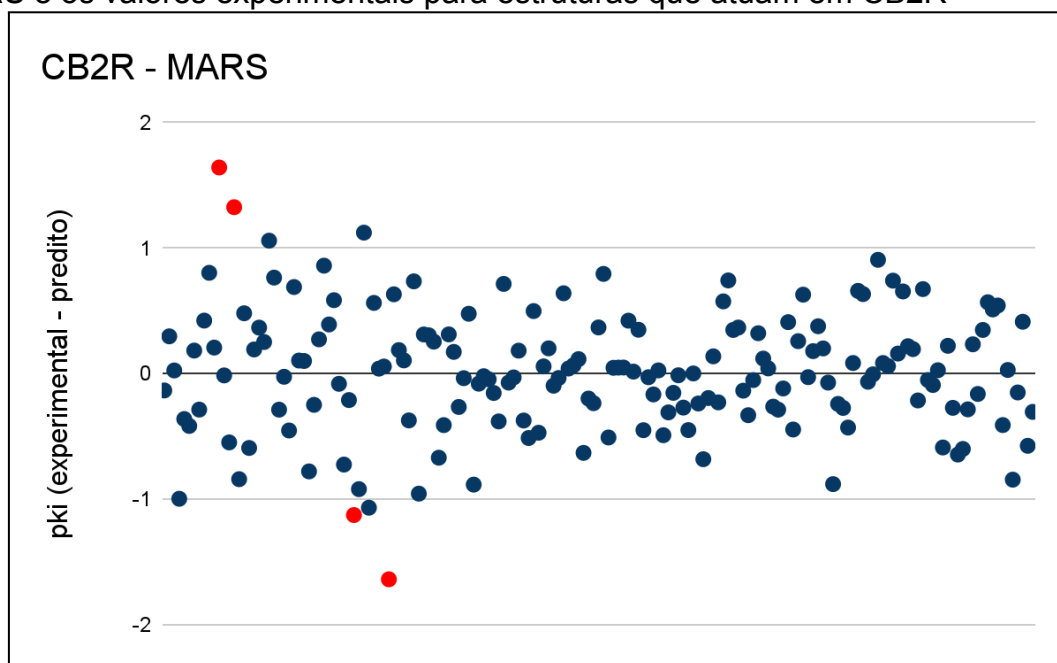
Fonte: Elaborada pela autora

Ao analisarmos, individualmente, cada caso percebeu-se que a exclusão de apenas 4 (1.3%) estruturas (representadas em vermelho no Figura 19) poderia resultar em um desempenho consideravelmente melhor. Logo, optamos por remover as estruturas CB1R_170, CB1R_173, CB1R_188 e CB1R_286 (Anexo B), no entanto, observamos que resultou em uma melhora pouco significativa e, sem a redução do sobreajuste.

Realizamos a mesma análise de dispersão entre os valores da diferença entre o *pKi* experimental e o predito para as estruturas que atuam no receptor CB2R.

Verificou-se que aproximadamente 7 estruturas (4%) apresentavam diferença de pKi superior a 1.0, conforme podemos verificar na Figura 20:

Figura 20 - Gráfico de resíduos com os valores da diferença entre pKi preditos por MARS e os valores experimentais para estruturas que atuam em CB2R



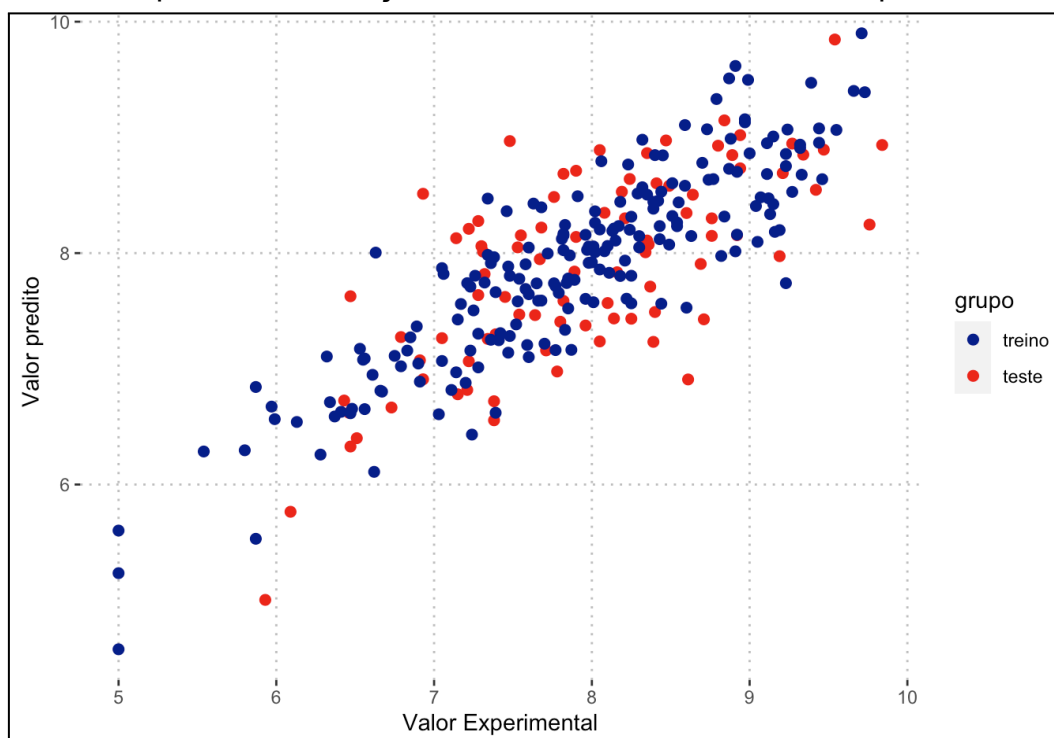
Legenda: Pontos em vermelho foram consideradas estruturas *outliers*.

Fonte: Elaborada pela autora

Verificando caso a caso decidiu-se excluir as estruturas CB2R_28, CB2R_33, CB2R_41 e CB2R_134 (Anexo C), que foram identificadas como *outliers* (representadas em vermelho na Figura 20). Entretanto, mesmo após a exclusão dessas estruturas, notou-se uma melhora pouco significativa na performance do modelo, sem redução do *overfitting*, conforme pode ser analisado na Tabela 08, coluna "excluindo *outliers*".

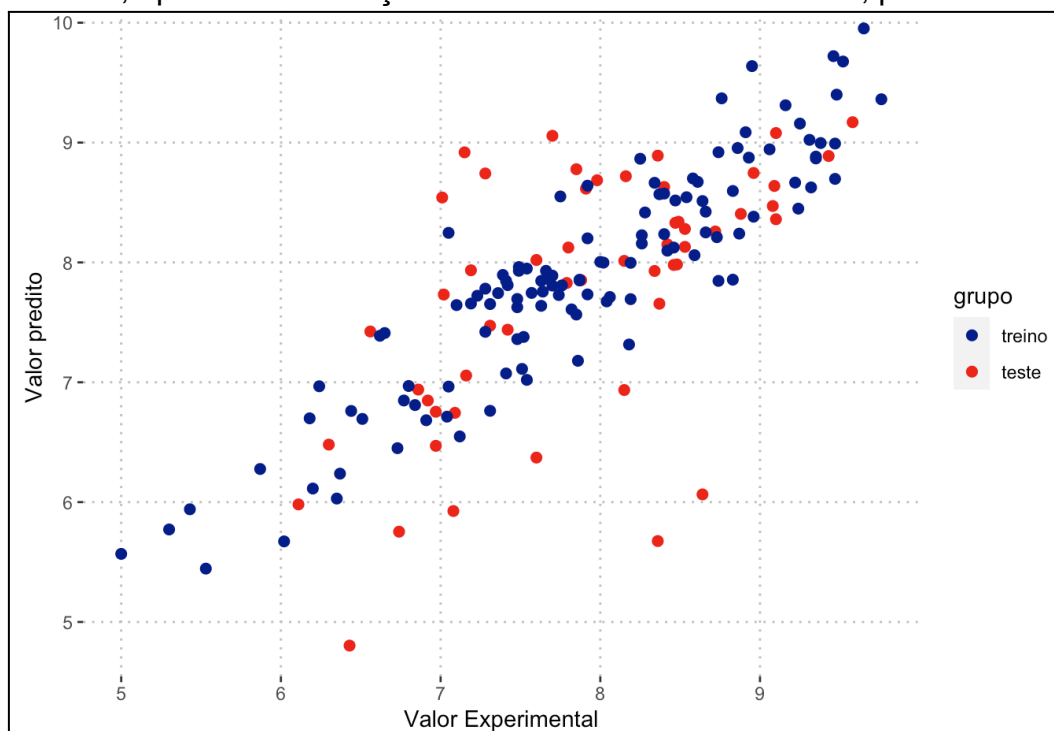
Com a finalidade de uma análise visual do que ocorreu com os modelos, foram gerados os gráficos comparativos entre valores treino e teste, utilizando o método MARS com todos os descritores, para os receptores CB1R e CB2R, representados nas Figuras 21 e 22, respectivamente.

Figura 21 - Comparativo entre as predições e experimental dos valores de pK_i obtidos nos conjuntos de treino e teste utilizando o método *MARS* com os descritores, após a clusterização dos dados e exclusão *outliers*, para CB1R.



Fonte: Elaborada pela autora

Figura 22 - Comparativo entre as predições e experimental dos valores de pK_i obtidos nos conjuntos de treino e teste utilizando o método *MARS* com os descritores, após a clusterização dos dados e exclusão *outliers*, para CB2R.



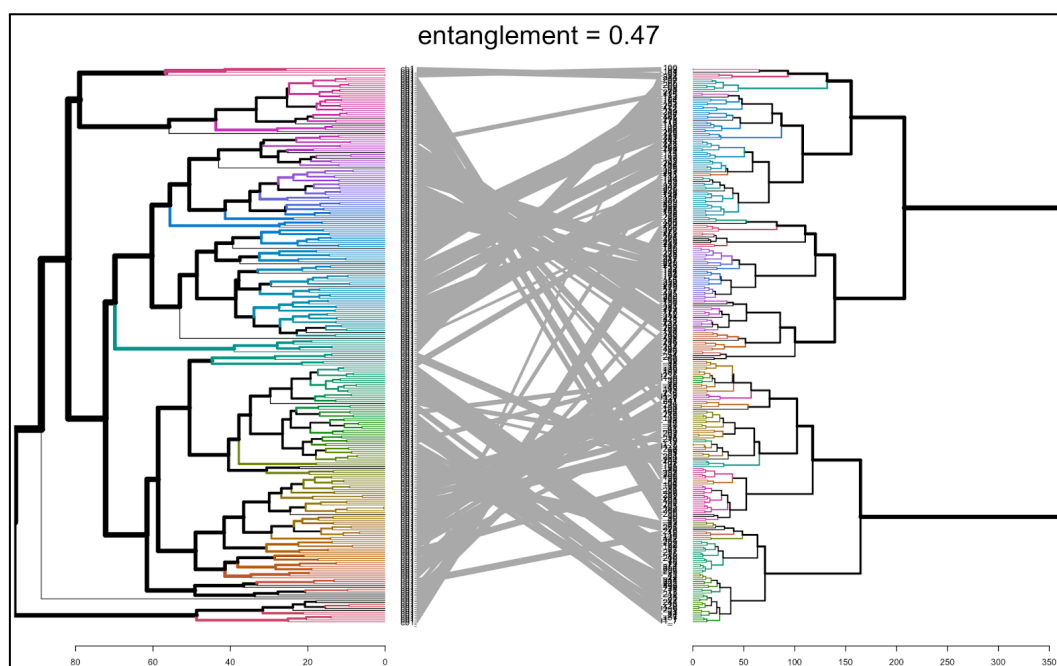
Fonte: Elaborada pela autora

É possível observar nas Figuras 21 e 22 a presença de outliers, o que pode ter contribuído para as divergências nas métricas entre o grupo teste e treino. A grande heterogeneidade das estruturas e uma possível falta de homogeneidade na divisão dos grupos, mesmo com as tentativas de tornar a divisão mais equilibrada, algumas divergências foram observadas nas métricas entre os grupos de teste e treino dos modelos. Essas divergências podem ser atribuídas, em parte, a uma falta de linearidade nos dados. Vale ressaltar que o modelo *MARS* é um modelo linear que pode ter suas limitações na captura de relações não lineares complexas presentes nos dados. Ao finalizar as tentativas de análises baseadas em *MARS* utilizando a linguagem R, verificamos que, com base nesse modelo, é necessário cautela na predição quanto aos canabinoides sintéticos atuando em CB1R e CB2R.

4.9 HCA e PCA

Para as análises não supervisionadas de Análise do Componente Principal (*PCA*) e Análise Hierárquica de Agrupamentos (*HCA*), as medidas de similaridades foram calculadas pela distância euclidiana e pelo método *Ward*. Pode ser verificado o emaranhamento entre o método euclidiano e o *Ward* nas Figuras 23 e 24 a seguir.

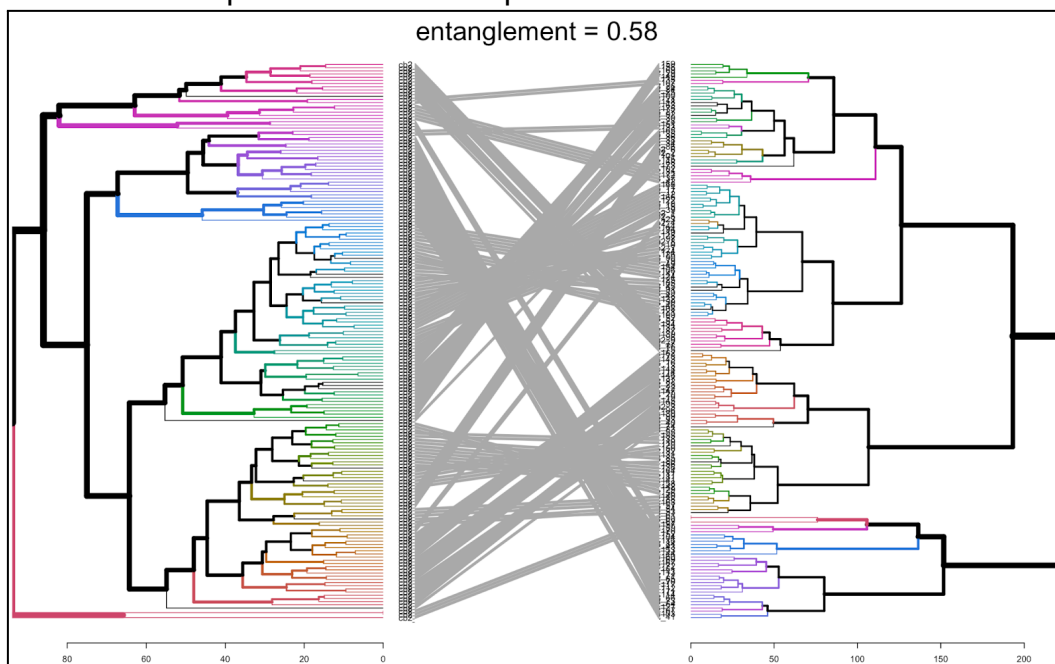
Figura 23 - Comparação entre dendrogramas pelos os métodos de distância euclidiana e *Ward* para as estruturas que atuam em CB1R



Legenda: Dendrograma à esquerda método de distância euclidiana e, à direita pelo método *Ward*.

Fonte: Elaborada pela autora.

Figura 24 - Comparação entre dendrogramas pelos os métodos de distância euclidiano e *Ward* para as estruturas que atuam em CB2R



Legenda: Dendrograma à esquerda método de distância euclidiano e, à direita pelo método *Ward*.

Fonte: Elaborada pela autora.

A análise dos dendrogramas gerados pelo método euclidiano e pelo método *Ward* (Figuras 23 e 24) revela que os agrupamentos possuem uma razoável similaridade, conforme evidenciado pelo baixo valor de emaranhamento (do inglês, *entanglement*) para o CB1R de 0.47 e para o CB2R de 0.58.

As medidas de similaridades foram calculadas pela distância euclidiana e pelo método *Ward* que visam minimizar a variação intra-grupos e maximizar a variação entre grupos, além de apresentar resultados sensíveis a *outliers* e ser aplicável a outras medidas de distância. O método *Ward* busca por partições que minimizam a perda associada a cada agrupamento e é calculada pela diferença entre a soma dos erros quadráticos de cada padrão e a média da partição em que está contido.

Foram realizados estudos utilizando os dados moleculares obtidos por meio dos descritores calculados, empregando os algoritmos *AGNES* e *DIANA*. Os resultados obtidos com esses algoritmos estão apresentados na Tabela 09:

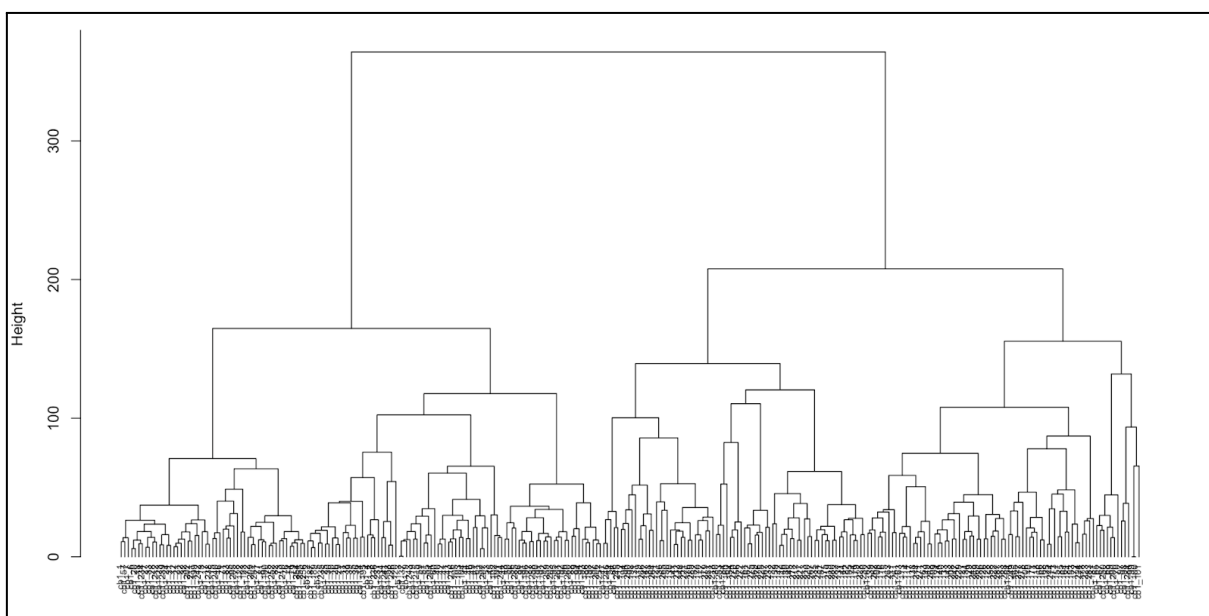
Tabela 09 - Coeficientes obtidos pelos métodos aglomerativo (*AGNES*) e divisivo (*DIANA*) na *HCA* para CB1R e CB2R.

Método	CB1R	CB2R
AGNES	0.840	0.824
DIANA	0.828	0.813

Fonte: Elaborada pela autora.

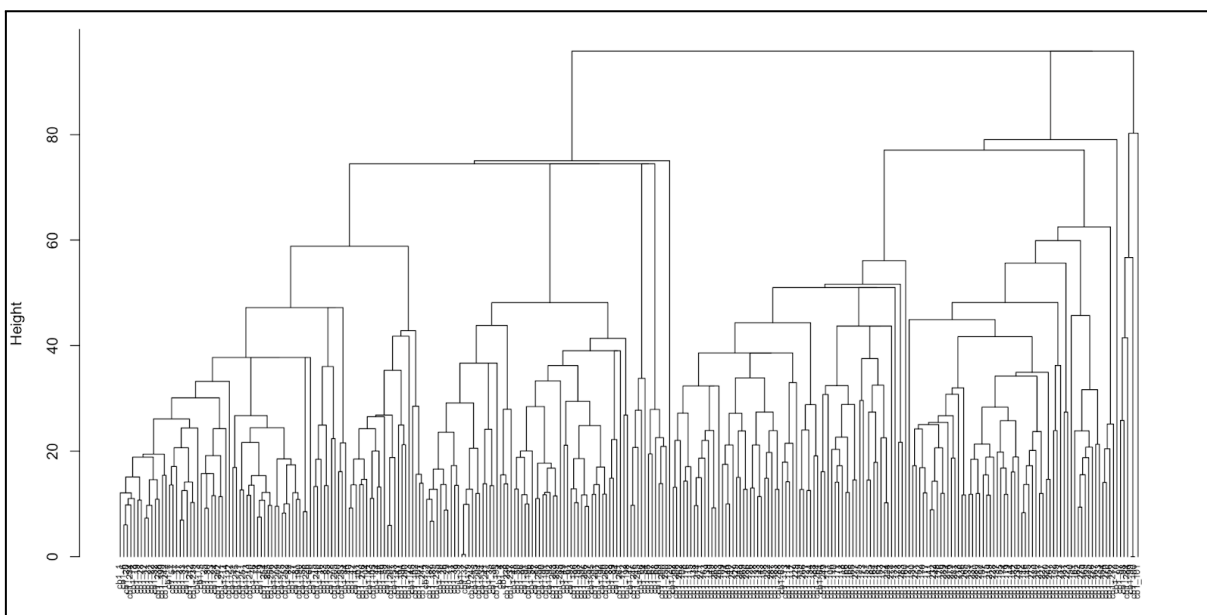
Em um estudo de *HCA*, os coeficientes aglomerativos podem variar entre 0 e 1, e valores mais elevados representam que estruturas muito claras foram identificadas. Todos os coeficientes aglomerativos apresentados na Tabela 09 são superiores a 0.8, indicando que houve uma separação considerada precisa. Isso pode ser confirmado pelas Figuras 25 a 28 a seguir.

Figura 25 - Análise por *HCA* para CB1R com todas as estruturas utilizando o método aglomerativo (*AGNES*)



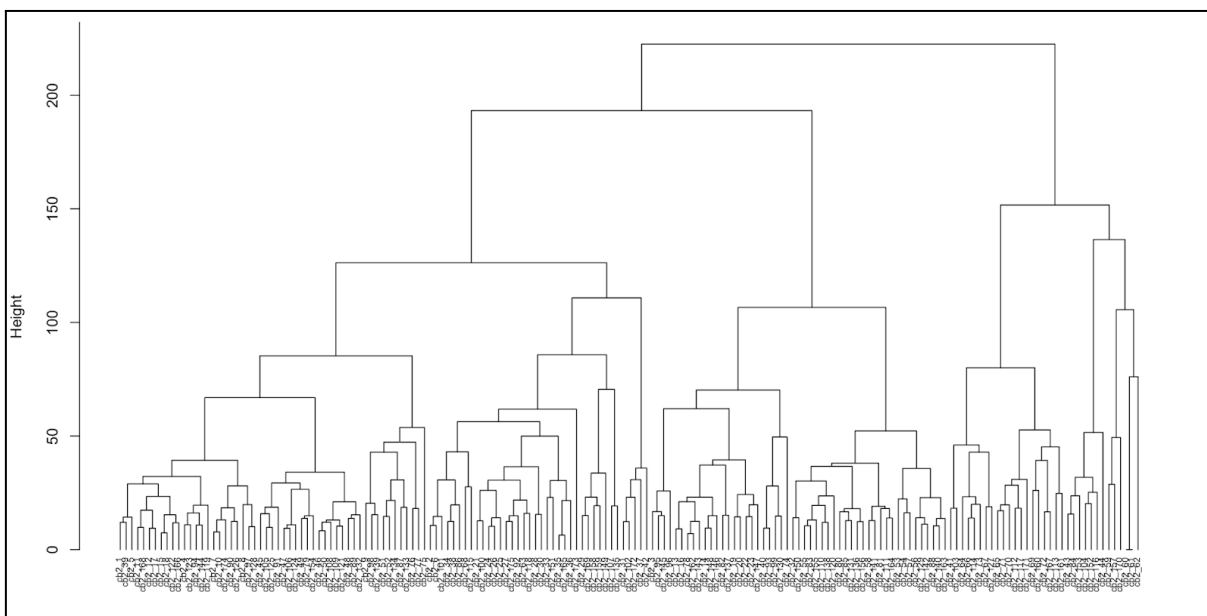
Fonte: Elaborada pela autora.

Figura 26 - Análise por HCA para CB1R com todas as estruturas utilizando o método divisivo (*DIANA*).



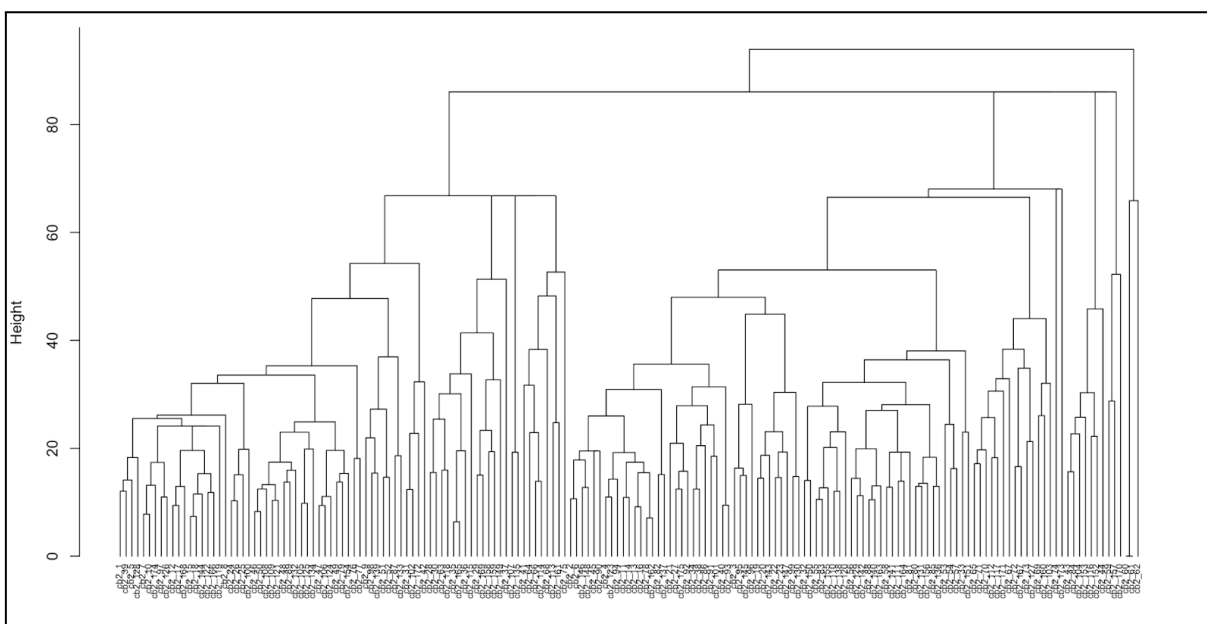
Fonte: Elaborada pela autora.

Figura 27 - Análise por HCA para CB2R com todas as estruturas utilizando o método aglomerativo (*AGNES*).



Fonte: Elaborada pela autora.

Figura 28 - Análise por HCA para CB2R com todas as estruturas utilizando o método divisivo (*DIANA*).



Fonte: Elaborada pela autora.

Ao examinar as Figuras (Figuras 25-28), pode-se observar que houve uma boa separação dos *clusters* e, que, tanto para CB1R quanto para CB2R, apresentaram melhor resultado quando foi utilizado método aglomerativo (AGNES). Dessa forma, é possível realizar uma melhor separação considerando inicialmente cada objeto um *cluster*, depois os dois *clusters* semelhantes são combinados em um novo cluster maior até atingir um platô.

Diante de uma divisão satisfatória, é necessário determinar o número ideal de *clusters* a serem utilizados. Nesse sentido, foram empregados o método *Elbow* e *Silhouette*, cujos resultados estão registrados na Tabela 10.

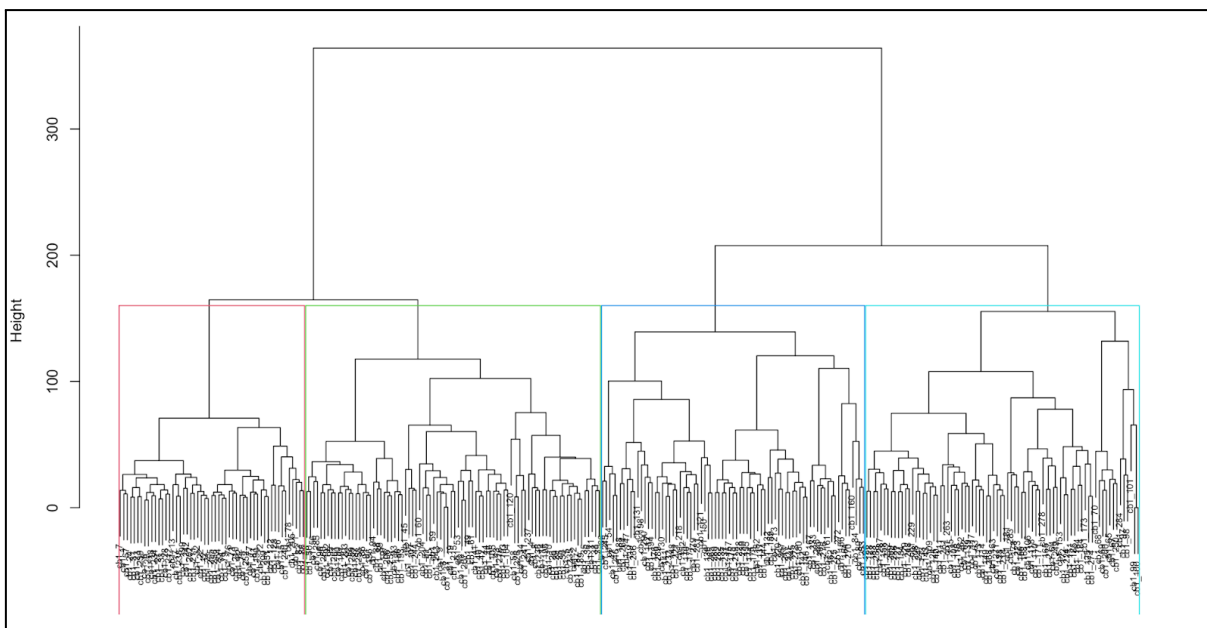
Tabela 10 - Número *cluster* sugerido para CB1R e CB2R, segundo o método *Elbow* e *Silhouette*

Método	CB1R	CB2R
<i>Elbow</i>	4	4
<i>Silhouette</i>	4	4

Fonte: Elaborada pela autora.

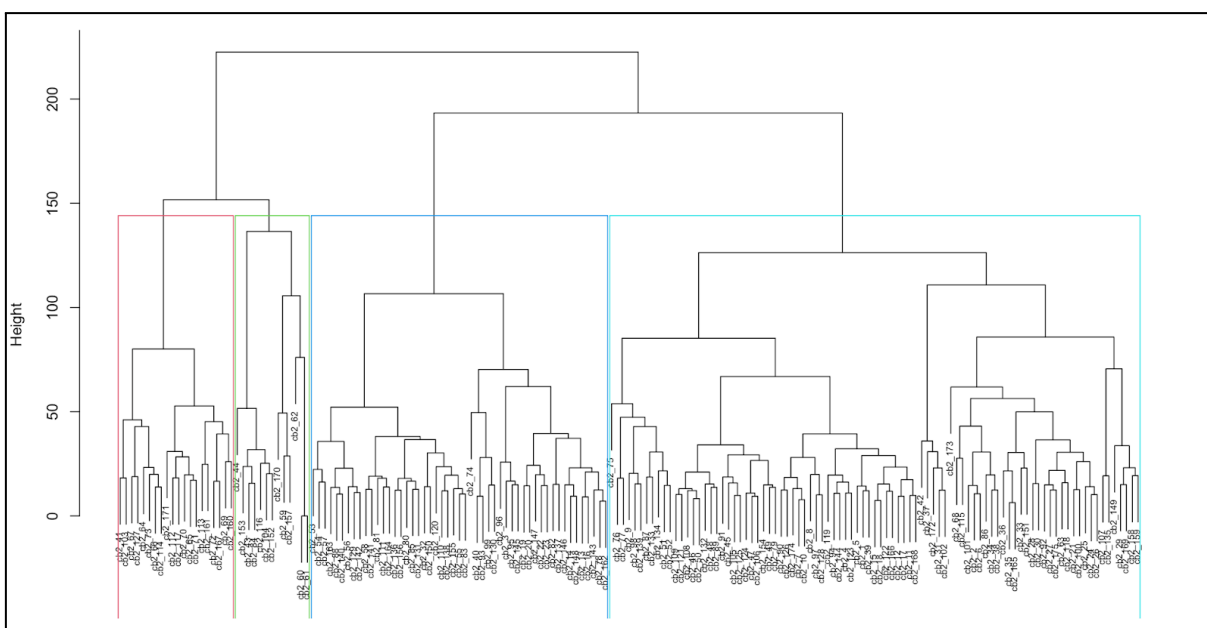
Na Tabela 10, é possível verificar que os dois métodos sugestivos corroboram que 4 *clusters* possibilitam uma boa separação dos dados. Portanto, de acordo com a sugestão apresentada, os dendrogramas podem ser divididos conforme ilustrado na Figura 29 e 30.

Figura 29 - Divisão de *clusters* obtida HCA para CB1R conforme sugeridos pelos os métodos *Elbow* e *Silhouette*.



Fonte: Elaborada pela autora.

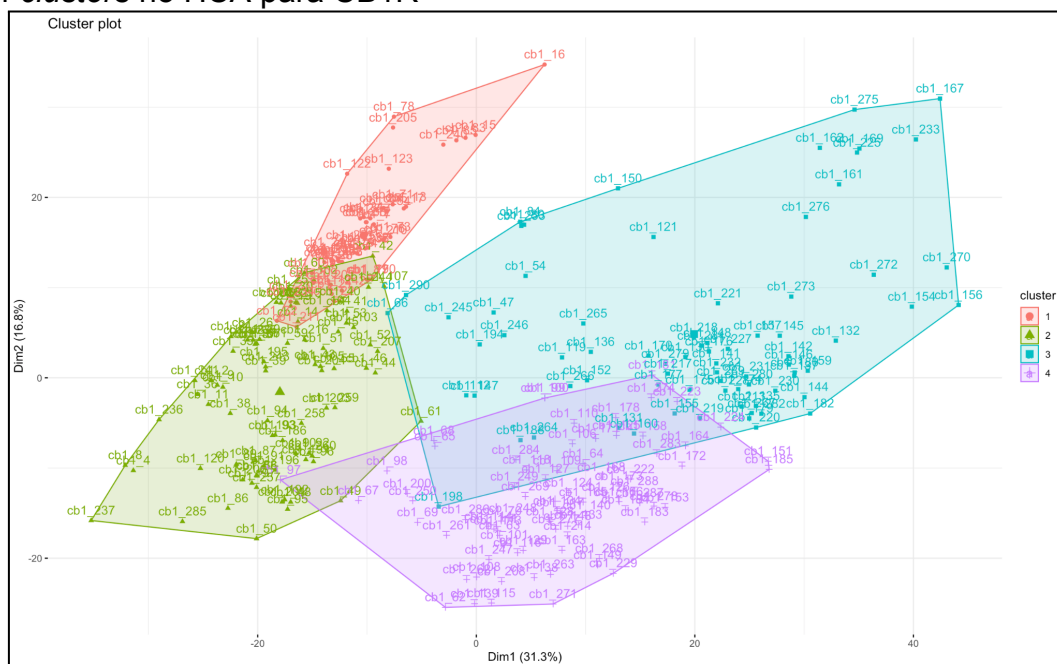
Figura 30 - Divisão de *clusters* obtida HCA para CB2R conforme sugeridos pelos os métodos *Elbow* e *Silhouette*.



Fonte: Elaborada pela autora.

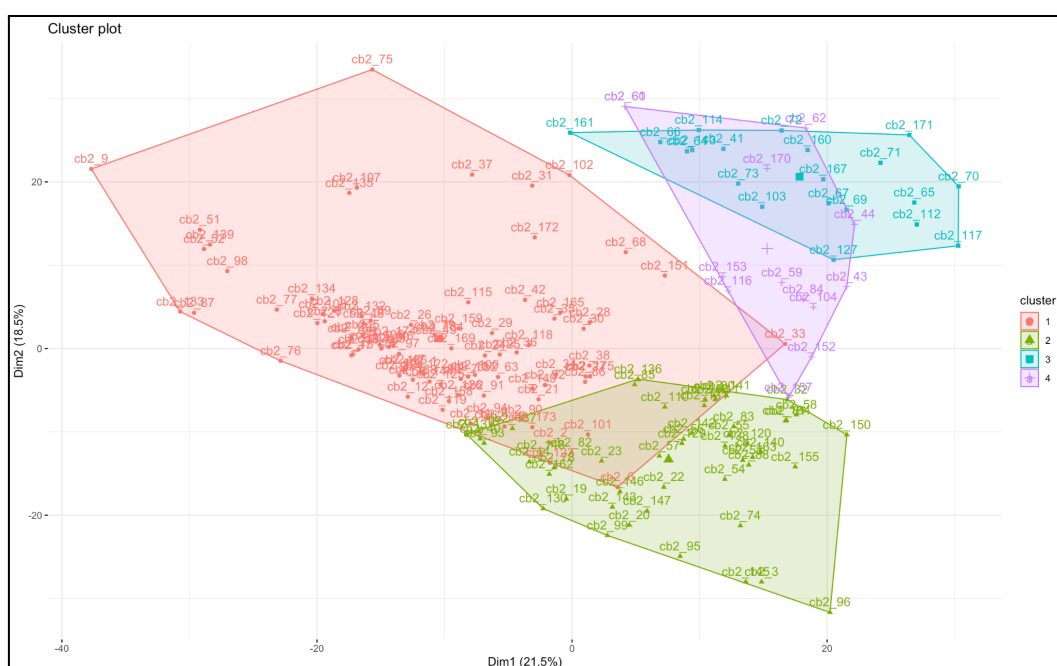
A aplicação do método da Análise do Componente Principal (*PCA*) corroborou com o *HCA*, confirmando que uma divisão em 4 grupos, como pode ser observado nas Figuras a seguir.

Figura 31 - Representação gráfica a partir das componentes principais (PC1 e PC2) dos 4 *clusters* no *HCA* para CB1R



Fonte: Elaborada pela autora.

Figura 32 - Representação gráfica a partir das componentes principais (PC1 e PC2) dos 4 *clusters* no *HCA* para CB2R



Fonte: Elaborada pela autora.

Foi realizado um estudo utilizando a Análise do Componente Principal (*PCA*) e a Análise Hierárquica de Agrupamentos (*HCA*), cujos resultados mostraram uma eficiente separação dos dados. Para definir o número ideal de *clusters*, foram aplicados os métodos *Elbow* e *Silhouette*, que confirmaram que 4 *clusters* possibilitam uma boa separação dos dados, o que foi corroborado tanto pelo *HCA* quanto pelo *PCA*. Além disso, os algoritmos *AGNES* e *DIANA* foram empregados para avaliar a precisão da separação dos agrupamentos, evidenciando que o método aglomerativo *AGNES* foi mais efetivo para ambos receptores. Dessa forma, pode-se afirmar que o uso de análise não supervisionadas permitiu uma separação precisa dos dados e pode ser uma abordagem útil em análises futuras.

5. Conclusões

Neste trabalho foram apresentados e discutidos os resultados obtidos pela aplicação de quimioinformática na obtenção de modelos capazes de reconhecer os perfis estruturais de canabinoides sintéticos e sua interação com os receptores canabinoides CB1R e CB2R. Para tanto, empregou-se técnicas de *Machine Learning* para construir modelos classificatórios robustos e foi empregada a curadoria dos dados para assegurar a acurácia dos cálculos dos descritores físico-químicos e das previsões de propriedades dos compostos. Mostrou-se importante a seleção criteriosa de descritores moleculares para garantir precisão do modelo estatístico.

O modelo de classificação baseado em árvore de decisão apresentou excelente desempenho na distinção entre as estruturas de canabinoides sintéticas que interagem com CB1R e CB2R. A redução da multicolinearidade demonstrou que *Forward Selection* e a *Backward Elimination* podem ser abordagens eficazes para obter modelos lineares simples e facilmente interpretáveis. O algoritmo *Random Forest* mostrou-se computacionalmente eficiente, robustez à não-linearidade, proporcionando resultados confiáveis e interpretação viável. O *Random Forest* se destacou como o melhor modelo não linear, evidenciando a presença de não linearidade nos dados e a capacidade dos descritores em fornecer um nível de informação relevante para a predição. Por outro lado, é importante ressaltar que o algoritmo *MARS*, deve ser usado com cautela na predição de canabinoides sintéticos atuando em CB1R e CB2R. Por fim, as técnicas de Análise do Componente Principal (*PCA*) e a Análise Hierárquica de Agrupamentos (*HCA*) permitiram uma separação precisa dos dados e podem ser abordagens úteis em análises futuras.

O objetivo primordial deste estudo foi contribuir para o avanço do conhecimento científico relacionado à interação dos canabinoides sintéticos com os receptores canabinoides (CB1R e CB2R). Além disso, os modelos gerados neste estudo possuem potencial de serem aplicados em estudos de triagem virtual e otimização de fármacos candidatos a medicamentos que possuam potencial atividade no sistema canabinóide.

REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, 2(4), 433-459. 2010.

DOI: 10.1002/wics.101

ADAMS, R.; PEASE, D. C.; CLARK, J. H. Isolation of Cannabinol, Cannabidiol and Quebrachitol from Red Oil of Minnesota Wild Hemp. **Journal of the American Chemical Society**, v. 62(8), p. 2194–2196. 1940.

DOI:10.1021/ja01865a080.

ALVES, V. M.; BRAGA, R. C.; MURATOVA, E. N. ; ANDRADE, C. H.

Quimioinformática: uma introdução. **Química Nova**. Vol. 41, No. 2. 2018.

DOI: 10.21577/0100-4042.20170145

AMIT, Y.; GEMAN, D. Shape Quantization and Recognition with Randomized Trees. **Neural Computation**. v. 9 (7), p. 1545–1588. 1997.

DOI:10.1162/neco.1997.9.7.1545

AN, D.; PEIGNEUR, S.; HENDRICKX, L. A.; TYTGAT, J. Targeting Cannabinoid Receptors: Current Status and Prospects of Natural Products. **International Journal of Molecular Sciences**. v. 21(14) . 2020.

DOI: 10.3390/ijms21145064

ARRUDA, A. C. S. **Ampliação e Aplicação Do Método Semi-empírico Topológico (let) Em Modelos Qsrr/Qspr/Qsar Para Compostos Alifáticos Halogenados e Cicloalcanos**. 2018. Tese(Doutorado em Química Analítica) - Universidade Federal de Santa Catarina. Florianópolis, 2008.

BENTO, A. P.; HERSEY, A.; FÉLIX, E.; LANDRUM, G.; GAULTON, A.; ATKINSON, F.; BELLIS, L. J.; DE VEIJ, M; LEACH, A. R. . An open source chemical structure curation pipeline using RDKit. **Journal of Cheminformatics**. v. 12(1). 2020.

DOI: 10.1186/s13321-020-00456-1

BOHACEK, R. S.; McMARTIN, C.; GUIDA, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. **Med Res Rev.** v. 16(1): 3-50 .1996.

DOI: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6

BREIMAN, L; **Using adaptive bagging to debias regressions.** Technical Report 547. Statistics Department. University of California at Berkeley. California. 1999.

DOI:10.1162/neco.1997.9.7.1545

BREIMAN, L. Random Forests. **Machine Learning.** v. 45, 5–32. 2001.

DOI: 10.1023/A:1010933404324

CABRAL, G. A.; FERREIRA, G. A.; JAMERSON, M. J. **Endocannabinoids and the Immune System in Health and Disease.** Handbook of experimental pharmacology. Ed. Pertwee. Springer International Publishing. 2015.

CAVALCANTI, A. B. S.; BARROS, R. P. C.; COSTA, V. C. De O.; SILVA, M. S. da; TAVARES, J. F.; SCOTTI, L. e SCOTTI, M. T. Computer-Aided Chemotaxonomy and Bioprospecting Study of Diterpenes of the Lamiaceae Family. **Molecules.** v. 24(21), p. 3908. 2019.

DOI: 10.3390/molecules24213908

ChemAxon.MarvinSketch version 20.15.0. ChemAxon Ltd., Budapest, Hungary, 2020.

CONSONNI, V.; TODESCHINI, R.; PAVAN, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. **Journal of Chemical Information and Computer Sciences.** v. 42, 3, 682–692. 2002.

DOI: 10.1021/ci015504a

DE VILLE, B. Decision trees. **Wiley Interdisciplinary Reviews: Computational Statistics.** v. 5(6), 448–455. 2013.

DOI:10.1002/wics.1278

DIETTERICH, T.G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. **Machine Learning** V. 40, 139–157. 2000.

DOI: 10.1023/A:1007607513941

DUDEK, A. Z.; ARODZ, T.; GÁLVEZ, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. **Combinatorial chemistry & high throughput screening**. v. 9(3), 213–228. 2006. DOI:

10.2174/138620706776055539

ENGEL, T.; GASTEIGER, J. **Applied Chemoinformatics - Achievements and Future Opportunities**. Wiley-VCH Verlag GmbH & Co. KGaA. 2018. Weinheim. Alemanha.

ISBN: 978-3-527-80654-6

ERIKSSON, L.; JOHANSSON, E.; KETTANEH-WOLD, N.; TRYGG, J.; WIKSTRÖM, C., WOLD, S. Multi-and megavariate data analysis: principles and applications. **Umetrics AB**. 2008.

ET-TALEBY, A.; BOUSSETTA, M.; BENSLIMANE, M. . Faults Detection for Photovoltaic Field Based on K-Means, Elbow, and Average Silhouette Techniques through the Segmentation of a Thermal Image. **International Journal of Photoenergy**.(2020)

DOI: 10.1155/2020/6617597

European Monitoring Centre for Drugs and Drug Addiction European Drug Report 2017: Trends and Developments. Publications Office of the European Union, Luxembourg. 2017.

FERREIRA, M.M. C.; ANTUNES A.M.; MELGO M.S . e VOLPE P. L.O. “Quimiometria I: Calibração Multivariada, um Tutorial”. **Revista Química Nova**. 1999.

FLORESTA,G.; APIRAKKAN O.; RESCIFINA A.; ABBATE V. Discovery of High-Affinity Cannabinoid Receptors Ligands through a 3D-QSAR Ushered by Scaffold-Hopping Analysis. **Molecules**.v. 23(9):2183. 2018.

DOI: 10.3390/molecules23092183

FRIEDMAN, J. H. Multivariate Adaptive Regression Splines. **The Annals of Statistics**. v 19(1), 1–67. 1991.

DOI: 10.1214/aos/1176347963

GALIÈGUE, S.; MARY, S.; MARCHAND, J.; DUSSOSSOY, D.; CARRIÈRE, D.; CARAYON, P.; BOUABOULA, M.; SHIRE, D.; LE FUR, G.; CASELLAS, P.

Expression of central and peripheral cannabinoid receptors in human immune tissues and leukocyte subpopulations. **European journal of biochemistry**. v. 232(1), 54–61. 1995

DOI: 10.1111/j.1432-1033.1995.tb20780.x

GASTEIGER, J.; FUNATSU, K. Chemoinformatics – An Important Scientific Discipline. **Journal of Computer Chemistry**. v., 5(2), 53–58. 2006.

DOI:10.2477/jccj.5.53

GENUER, R.; POGGI, J. M.; TULEAU-MALOT, C. Variable Selection using Random Forests. **Pattern Recognition Letters**. v. 31(14), 2225-2236. 2010.

GILMOUR, S. G.. The Interpretation of Mallows's C_p -Statistic. **The Statistician**. v. 45(1). 1996

DOI:10.2307/2348411

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning-Data Mining, Inference, and Prediction**. 2 Ed. Springer Series in Statistics. Nova Iorque. 2009.

DOI: 10.1007/978-0-387-84858-7

HO, T. K. . Random decision forests. **Proceedings of 3rd International Conference on Document Analysis and Recognition**. 1995.

HUA, T.; Li, X.; Wu, L.; ILIOPOULOS-TSOUSOUVAS, C.; WANG Y.; WU, M.; SHEN, L.; BRUST, C.A.; NIKAS, S. P.; SONG, F.; SONG, X.; Yuan, S.; SUN, Q.; WU, Y.; JIANG, S.; GRIM, T.W.; BENCHAMA, O.; STAHL, E.L.; ZVONOK, N.; ZHAO, S.; BOHN, L.M.; MAKRIYANNIS, A.; LIU, Z.J. Activation and Signaling Mechanism Revealed by Cannabinoid Receptor-GiComplex Structures. **Cell**. v. 180, 655–665. 2020.

DOI: 10.1016/j.cell.2020.01.008

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. 2 Ed .Springer Series in Statistics. Nova Iorque. 2013.

JOLLIFFE, I. T. **Principal Component Analysis for Special Types of Data**. In: **Principal Component Analysis**. Springer Series in Statistics. Nova Iorque. 2002.
DOI: 10.1007/0-387-22440-8_13

KATRITZKY, A. R.; GORDEEVA, E. V. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. **Journal of Chemical Information and Computing Science**. v. 33(6), 835–857.1993.
DOI: 10.1021/ci00016a005

LU, D.; POTTER, D. E. **Handbook of Cannabis and Related Pathologies**. “Chapter 58 - Cannabinoids and the Cannabinoid Receptors: An Overview.” Academic Press. São Diego. 2017.
DOI: 10.1016/B978-0-12-800756-3.00068-5.

LU, H.C.; MACKIE, K. Review of the Endocannabinoid System. **Biological psychiatry. Cognitive neuroscience and neuroimaging**. v. 6 (6). 2021.
DOI: 10.1016/j.bpsc.2020.07.016.

MAHESH, B. Machine Learning Algorithms - A Review . **International Journal of Science and Research (IJSR)**. v. 9. 2020.
DOI:10.21275/ART20203995

MATSUDA, L. A.; LOLAIT, S. J.; BROWNSTEIN, M. J.; YOUNG, A. C.; BONNER, T. I. Structure of a cannabinoid receptor and functional expression of the cloned cDNA. **Nature**. v. 346(6284), 561–564. 1990.
DOI: 10.1038/346561a0

Minitab, LLC Support. Disponível em:

<https://support.minitab.com/pt-br/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/model-assumptions/multicollinearity-in-regression/> Acesso 31 dezembro 2021.

MEDINA-FRANCO J.L.; CHÁVEZ-HERNÁNDEZ A.L.; LÓPEZ-LÓPEZ E.;
SALDÍVAR-GONZÁLEZ F.I.. Chemical Multiverse: An Expanded View of Chemical
Space. **Molecular Informatics**. v. 41(11). 2022

DOI: 10.1002/minf.202200116.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. . **Introduction to Linear
Regression Analysis** 5 Ed. John Wiley & Sons. 2012.

MUNRO, S.; THOMAS, K. L.; ABU-SHAAR, M. Molecular characterization of a
peripheral receptor for cannabinoids. **Nature**. v 365(6441), 61–65. 1993.

DOI: 10.1038/365061a0

NEATH, A. A.; CAVANAUGH, J. E. The Bayesian information criterion: background,
derivation, and applications. **WIREs Computational Statistics** v. 4(2). 2011.

DOI:10.1002/wics.199

NEVES B. J.; BRAGA R. C.; MELO-FILHO C. C.; MOREIRA-FILHO J.; MURATOV E.
N.; ANDRADE, C.H. QSAR-Based Virtual Screening: Advances and Applications in
Drug Discovery. **Frontiers in Pharmacology**. V. 9 (1275). 2018

DOI: 10.3389/fphar.2018.01275.

PEREIRA, F. da S. **Estudo QSAR da artemisinina e derivados com atividade
antimalárica**. 2016. Tese (Doutorado em Ciências) - Instituto de Química -
Universidade de Campinas. Campinas, 2016

PETERS, H. **A Brief History of Four Millennia (B.C. 2000—A.D. 1974)**. In
Marihuana and Medicine. Editado por NAHAS, G. G.; SUTIN, K. M.; HARVEY, D.;
AGURELL, S.; PACE, N. ; CANCRO R. MLA. Humana Press. Totowa, Nova Jersey
1999.

DOI: 10.1007/978-1-59259-710-9_1.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning
strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and
Knowledge Discovery**, e1301. 2019.

DOI: 10.1002/widm.1301

PROVOST, F.; FAWCETT, T. .Analysis and visualization of classifier performance:
Comparison under imprecise class and cost distribution. **Proceedings of the 3rd**

Intl. Conf, on Knowledge Discovery and Data Mining (KDD-97). Menlo Park, CA: AAAI Press, pp. 43-48. 1997

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em www.R-project.org/. Acesso: 31 dezembro 2021.

RODRÍGUEZ-PÉREZ, R; BAJORATH, J. Explainable Machine Learning for Property Predictions in Compound Optimization. **Journal of Medicinal Chemistry**. v. 64 (24), 17744-17752. 2021.

DOI: 10.1021/acs.jmedchem.1c01789

SANTOS, A. M. N. **Análise toxicológica de canabinoides sintéticos em contexto forense**. 2018. Dissertação (Mestrado). Faculdade Ciências da Saúde Universidade Fernando Pessoa Porto. Porto, 2018

SCHICHO, R.; STORR, M. Alternative Targets Within the Endocannabinoid System for Future Treatment of Gastrointestinal Diseases. **Canadian Journal of Gastroenterology**. v. 25(7), 377–383. 2011.

DOI: 10.1155/2011/953975

SCHNEIDER, G. Automating drug discovery. **Nature Reviews Drug Discovery**. v. 17(2), 97–113. 2018.

DOI: 10.1038/nrd.2017.232

SHAHBAZI F.; GRANDI, V.; BANERJEE, A; TRANT, J.F. Cannabinoids and Cannabinoid Receptors: The Story so Far. **iScience**. v. 23 (7): 101301. 2020.

DOI: 10.1016/j.isci.2020.101301

SILVA, T. M. B. **Estudo químico-teórico na diferenciação da psicoatividade e psicoinatividade em canabinoides biologicamente ativos e seus metabólitos**. 2009. Tese (Doutorado em Química e Biotecnologia) - Instituto de Química e Biotecnologia, Universidade Federal de Alagoas. Maceió, 2009.

SONG Y.Y. ; LU, Y. Decision tree methods: applications for classification and prediction. **Shanghai Archives of Psychiatry** . vl. 27, No. 2. 2015.

DOI: 10.11919/j.issn.1002-0829.215044

SPILLER K.J.; BI G.H.; HE Y.; GALAJ E.; GARDNER E.L.; Xi Z.X. Cannabinoid CB1 and CB2 receptor mechanisms underlie cannabis reward and aversion in rats.

British Journal of Pharmacology. v. 176(9):1268-1281. 2019

DOI: 10.1111/bph.14625.

STELLA, N. Cannabinoid and cannabinoid-like receptors in microglia, astrocytes, and astrocytomas. **Glia**. v. 58, 1017–1030. 2010.

DOI: 10.1002/glia.20983

SVETNIK, V.; LIAW, A.; TONG, C.; CULBERSON, J. C.; SHERIDAN, R. P.; FEUSTON, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. **Journal of Chemical Information and Computer Sciences**. v. 43(6), 1947–1958. 2003.

DOI: 10.1021/ci034160g

TANAKA, M.; SACKETT, S.; ZHANG, Y. Endocannabinoid Modulation of Microglial Phenotypes in Neuropathology. **Frontiers in neurology**. 11, 87–87. 2020.

DOI: 10.3389/fneur.2020.00087

TODESCHINI, R.; CONSONNI, V. **Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing**. John Wiley & Sons. 2009.

TODESCHINI, R.; CONSONNI, V.; XIANG, H.; HOLLIDAY, J.; BUSCEMA, M.; WILLETT, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. **Journal of Chemical Information and Modeling**. v. 52(11), 2884–2901. 2012

DOI:10.1021/ci300261r

TROPSHA, A.; GRAMATICA, P.; GOMBAR, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. **QSAR Combinatorial Science**, 22(1), 69-77. 2003.

TROPSHA, A. Best Practices for QSAR Model Development, Validation, and Exploitation. **Molecular Informatics**. v. 29 (6-7), 476–488. 2010.

DOI: 10.1002/minf.201000061

United Nations Office on Drugs and Crime (UNODC). **Synthetic cannabinoids in herbal products**. 2010.

VALE, M. N. **Avaliação de métodos e desenvolvimento de aplicativo para análise de grupos**. 2005. Dissertação (Mestrado) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica. Rio de Janeiro, 2005.

WANG, Y.; LI, Y.; QIAO, C. ; LIU, X.; HAO, M.; SHUGART, Y. Y.; XIONG, M.; JIN, L. Nuclear Norm Clustering: a promising alternative method for clustering tasks. **Scientific Reports - Nature**. v. 8, 10873. 2018.
<https://doi.org/10.1038/s41598-018-29246-4>

WEIS, W. I.; KOBILKA, B. K. The Molecular Basis of G Protein-Coupled Receptor Activation. **Annual review of biochemistry**, 87, 897–919. 2018.
DOI: 10.1146/annurev-biochem-060614-033910

WU, W.; MASSART, D.L.; JONG S. de. The kernel PCA algorithms for wide data. Part I: Theory and algorithms, **Chemometrics and Intelligent Laboratory Systems**. v. 36(2), 165–172. 1997.
DOI: 10.1016/S0169-7439(97)00010-5.

YAP, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. **Journal of Computational Chemistry**. v. 32, 1466–1474. 2011.

ZOU, S.; KUMAR, U. Cannabinoid Receptors and the Endocannabinoid System: Signaling and Function in the Central Nervous System. **International Journal of Molecular Sciences**. v. 19(3), 833. 2018.
DOI: 10.3390/ijms19030833

Nome	Nomenclatura	<i>pKi</i> Exp_CB1	<i>pKi</i> Exp_CB2
fl_46	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(OC)c4ccccc43</chem>	6.32	07.01
fl_57	<chem>O=C(c1c2ccccc2n(c1CCCC)CCCC)c3ccc(OC)c4ccccc43</chem>	6.85	6.51
fl_68	<chem>O=C(c1c2ccccc2n(c1CCCCC)CCCC)c3ccc(OC)c4ccccc43</chem>	6.34	6.92
fl_79	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(OC)c4ccccc43</chem>	7.45	7.75
fl_97	<chem>Clc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2C)cc1</chem>	5.8	5.43
fl_98	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4ccc(cc43)C</chem>	6.46	7.79
fl_100	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc43)C</chem>	6.91	7.85
fl_102	<chem>Fc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	7.39	7.48
fl_103	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccc(OC)cc3</chem>	5.97	6.35
fl_104	<chem>Fc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2)c1</chem>	7.14	07.04
fl_105	<chem>Clc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2C</chem>	7.89	7.60
fl_107	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(OC)c4ccccc43</chem>	7.2	7.49
fl_108	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccc(C)cc3</chem>	6.75	6.24
fl_109	<chem>Clc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2)c1</chem>	7.42	6.97
fl_111	<chem>Clc1ccc(C(=O)c2c3ccccc3n(CCCCC)c2C)c4ccccc14</chem>	08.05	8.64
fl_112	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3cccc4ccccc43</chem>	6.47	7.86
fl_114	<chem>FCCCCn1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31</chem>	7.53	9.22
fl_115	<chem>Fc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	7.64	7.41
fl_116	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3ccc(OC)c4ccccc43</chem>	7.47	7.88
fl_117	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccccc3OC</chem>	7.6	07.09
fl_118	<chem>O=C(Oc1cccc2ccccc12)c3c4ccccc4n(n3)CCCC</chem>	8.44	8.58
fl_119	<chem>O=C(c1cc(n(CCCCC)c1)-c2ccc3ccccc3c2)c4cccc5ccccc54</chem>	6.48	6.77
fl_120	<chem>Clc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	8.1	8.15
fl_121	<chem>FCCCCn1cc(c2ccccc21)C(=O)Nc3ccc4ccccc4c3</chem>	6.63	6.65
fl_123	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2)c1)c3cccc4ccccc43</chem>	7.85	8.19
fl_124	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4ccc(OC)cc43</chem>	8.18	8.16
fl_125	<chem>O=C(c1cn(CCCCC)c(-c2ccc(OC)c2)c1)c3cccc4ccccc43</chem>	7.28	7.64
fl_126	<chem>O=C(c1cn(CCCC)c2ccccc21)c3ccc(OC)c4ccccc43</chem>	8.25	8.66
fl_127	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(c4ccccc43)CCC</chem>	9.19	8.96
fl_130	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3cccc4ccccc43</chem>	08.01	8.26
fl_131	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3cccc4ccccc43</chem>	9	8.59

Nome	Nomenclatura	pKi Exp_CB1	pKi Exp_CB2
fl_142	<chem>Brc1ccc(c2ccccc12)C(=O)c3cn(CCCCC)c4ccccc43</chem>	8.92	8.96
fl_146	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(c4ccccc43)CCC</chem>	7.59	08.02
fl_151	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccccc43</chem>	08.02	8.53
fl_152	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3cccc4ccccc43</chem>	7.66	8.37
fl_153	<chem>O=C(c1c2ccccc2n(c1C)CC)c3cccc4ccccc43</chem>	5.93	06.02
fl_154	<chem>O=C(c1cn(c2ccccc21)C)c3cccc4ccccc43</chem>	5.00	5.00
fl_155	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(OC)c4ccccc43</chem>	8.35	8.73
fl_156	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(CCCC)c4ccccc43</chem>	6.47	7.28
fl_157	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc43)CCC</chem>	7.28	7.92
fl_158	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(CCCC)c4ccccc43</chem>	7.38	8.19
fl_159	<chem>Brc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	6.79	7.57
fl_160	<chem>Fc1ccc(C(=O)c2c3ccccc3n(CCCCC)c2C)c4ccccc41</chem>	7.85	8.66
fl_161	<chem>O=C(c1c2ccccc2n(n1)CCCC)c3cccc4ccccc43</chem>	8.23	8.34
fl_162	<chem>Fc1ccc(Cn2cc(c3ccccc32)C(Oc4ccccc5ccccc45)=O)cc1</chem>	8.92	8.61
fl_163	<chem>O=C(NC12CC3CC(C1)CC(C2)C3)c4c5ccccc5n(CCCCC)c4</chem>	8.19	8.91
fl_164	<chem>FCCCCn1c2ccccc2c(n1)C(=O)NC34CC5CC(C3)CC(C4)C5</chem>	8.71	9.58
fl_166	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)C</chem>	9.11	9.35
fl_167	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)C)C(OC)=O)cc1</chem>	8.00	9.10
fl_168	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc43)CC</chem>	7.15	7.92
fl_169	<chem>F[C@@H](Cn1cc(C(=O)C2C(C2)C)(C)C)c3ccccc31)CCC</chem>	7.23	8.74
fl_170	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2)c1)c3cccc4ccccc43</chem>	7.96	8.15
fl_171	<chem>Clc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)c1</chem>	7.15	7.80
fl_172	<chem>FC(F)(F)c1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	6.66	7.28
fl_173	<chem>O=C(c1cn(CCCCC)c(-c2ccc(CC)cc2)c1)c3cccc4ccccc43</chem>	7.47	7.54
fl_174	<chem>O=C(c1cn(c(-c2ccccc2CC)c1)CCCC)c3cccc4ccccc43</chem>	7.77	8.47
fl_175	<chem>O=C(c1cn(c(-c2ccccc2C)c1)CCCC)c3cccc4ccccc43</chem>	8.25	8.40
fl_176	<chem>Clc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2)cc1</chem>	6.41	6.30
fl_177	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccc(C)cc3</chem>	6.13	5.87
fl_178	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccc(OC)c3</chem>	7.21	07.08
fl_179	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(c4ccccc43)CC</chem>	7.48	8.00
fl_180	<chem>Fc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2C</chem>	7.41	7.12

Nome	Nomenclatura	pKi Exp_CB1	pKi Exp_CB2
fl_181	<chem>O=C(OC)[C@H](NC(=O)c1c2ccccc2n(CCCCC)c1)C(C)C</chem>	7.82	7.85
fl_182	<chem>ClCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)C)C(=O)N</chem>	8.39	7.92
fl_183	<chem>FCCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)(C)C)C(=O)N</chem>	8.84	9.16
fl_184	<chem>FCCCCCn1c2ccccc2c(n1)C(=O)N[C@H](Cc3ccccc3)C(=O)N</chem>	6.90	7.76
fl_190	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(c4ccccc34)CC</chem>	8.82	9.38
fl_201	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4ccc(CC)cc34</chem>	08.08	8.42
fl_206	<chem>FC(F)(F)CCCCc1cc(O)c2c(OC([C@@H]3CC=C(C[C@@H]23)C)(C)C)c1</chem>	7.70	7.52
fl_212	<chem>O=C(c1c2ccccc2n(CCC)c1)c3cccc4ccc(CC)cc34</chem>	6.47	6.91
fl_213	<chem>O=C(Oc1cccc2cccn21)c3c4ccccc4n(CC5CCCC5)c3</chem>	9.66	9.47
fl_214	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3cccc4ccccc43</chem>	08.05	8.53
fl_215	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4ccc(CC)cc43</chem>	5.87	6.62
fl_216	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(CCCC)c4ccccc34</chem>	6.83	7.31
fl_217	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(OCC)c4ccccc43</chem>	8.34	7.98
fl_218	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccc(CC)cc43</chem>	7.55	8.25
fl_219	<chem>BrC1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	6.43	7.51
fl_220	<chem>Clc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	6.73	7.66
fl_221	<chem>Clc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	07.03	7.36
fl_223	<chem>Fc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	6.62	7.48
fl_224	<chem>Fc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	6.28	7.42
fl_225	<chem>O=C(c1c(n(c2ccccc21)C)C)c3cccc4ccccc43</chem>	5	5.30
fl_226	<chem>FCCCCCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc43)C</chem>	8.8	9.24
fl_227	<chem>F[C@@H](CCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc43)C)C</chem>	8.49	09.09
fl_228	<chem>FCCCCCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc34)CC</chem>	9.42	9.43
fl_229	<chem>FCCCCCn1c2ccccc2c(n1)C(=O)c3cccc4ccccc43</chem>	8.87	8.88
fl_230	<chem>FCCCCCn1cc(c2ccccc21)C(Oc3cccc4cccn43)=O</chem>	9.33	9.20
fl_231	<chem>Fc1ccccc1-n2c3ccccc3c(n2)C(Oc4ccc5ccccc5c4)=O</chem>	5	6.37
fl_232	<chem>FCCCCCn1c2ccccc2c(n1)C(Oc3cccc4ccccc34)=O</chem>	8.59	8.47
fl_233	<chem>FCCCCCn1cc(c2ccccc21)C(=O)NC34CC5CC(C3)CC(C4)C5</chem>	8.6	9.10
fl_234	<chem>O=C(NCc1ccccc1)c2c3ccccc3n(CCCCC)c2</chem>	7.28	6.73
fl_235	<chem>FCCCCCn1cc(c2ccccc21)C(=O)NCc3ccccc3</chem>	7.14	6.37
fl_236	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.23	8.46

Nome	Nomenclatura	pKi Exp_CB1	pKi Exp_CB2
fl_237	<chem>O=C(Nc1cccc1)c2cn(CCCCC)c3cccc32</chem>	6.79	6.56
fl_238	<chem>O=C(Nc1cccc2cccc21)c3cn(CCCCC)c4cccc43</chem>	8.74	7.66
fl_239	<chem>FCCCCCn1cc(c2cccc21)C(=O)Nc3cccc4cccc43</chem>	8.43	7.87
fl_240	<chem>ClCCCCn1cc(c2cccc21)C(=O)Nc3cccc4cccc43</chem>	8.43	7.87
fl_241	<chem>Fc1ccc(Cn2c3cccc3c(n2)C(=O)NC45CC6CC(C4)CC(C5)C6)cc1</chem>	8.97	9.76
fl_242	<chem>O=C(Nc1cccc2cccc21)c3c4cccc4n(n3)CCCC</chem>	8.41	8.46
fl_243	<chem>O=C(c1c2cccc2n(CCCCC)c1)c3cccc3OC</chem>	7.58	8.54
fl_244	<chem>O=C(c1c2cccc2n([C@H](CCCC)C)c1C)c3cccc4cccc43</chem>	7.32	8.40
fl_245	<chem>BrC1CCCC2CCCC(c12)C(=O)c3cn(CCCCC)c4cccc43</chem>	7.68	8.26
fl_246	<chem>O=C(c1cn(CCCCC#N)c2cccc21)c3cccc4cccc43</chem>	9.55	8.83
fl_247	<chem>O=C(c1c2cccc2n(CCCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.97	9.31
fl_249	<chem>FCCCCCn1cc(c2cccc21)C(=O)NC(c3cccc3)(C)C</chem>	7.3	7.10
fl_250	<chem>O=C(N)[C@@H](NC(=O)c1c2cccc2n(n1)CCCC)C(C)C</chem>	8.54	09.06
fl_252	<chem>Fc1ccc(Cn2c3cccc3c(n2)C(=O)N[C@@H](C(C)(C)C)C(OC)=O)cc1</chem>	8.94	9.92
fl_253	<chem>O=C(C1C(C1(C)C)(C)C)c2c3cccc3n(CC4CCOCC4)c2</chem>	7.6	9.25
fl_254	<chem>O=C(C12CC3CC(C1)CC(C2)C3)c4c5cccc5n(CCCCC)c4</chem>	7.48	8.76
fl_255	<chem>O=C(C1C(C1(C)C)(C)C)c2c3cccc3n(CCCCC)c2</chem>	7.25	8.83
fl_256	<chem>FCCCCCn1c(c(C(=O)C2C(C2(C)C)(C)C)c3cccc31)C</chem>	7.71	8.34
fl_257	<chem>FCCCCCn1c2cccc2c(n1)C(=O)C3C(C3(C)C)(C)C</chem>	7.76	9.35
fl_258	<chem>O=C(c1c2cccc2n(CCCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.26	7.49
fl_259	<chem>FC(F)(F)CCn1cc(C(=O)C2C(C2(C)C)(C)C)c3cccc31</chem>	7.36	09.08
fl_260	<chem>O=C(c1cn(CCCCCC)c(-c2cccc2)c1)c3cccc4cccc43</chem>	7.68	7.21
fl_261	<chem>O=C(c1cn(CCCC)c(-c2cccc2)c1)c3cccc4cccc43</chem>	7.22	7.82
fl_262	<chem>O=C(c1cn(CCCCC)c(-c2ccc(OC)cc2)c1)c3cccc4cccc43</chem>	6.55	7.39
fl_263	<chem>O=C(c1cn(CCCCC)c(-c2ccc(cc2)C)c1)c3cccc4cccc43</chem>	6.89	7.74
fl_264	<chem>Clc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)cc1</chem>	6.56	7.60
fl_265	<chem>O=C(c1cn(CCCCC)c(-c2cccc2OC)c1)c3cccc4cccc43</chem>	7.54	7.70
fl_266	<chem>Fc1cccc1-c2cc(cn2CCCC)C(=O)c3cccc4cccc43</chem>	8.11	8.48
fl_267	<chem>O=C(c1cc(n(CCCCC)c1)-c2cccc3cccc32)c4cccc5cccc54</chem>	7.39	7.31
fl_268	<chem>O=C(c1cn(CCCCC)c(-c2cccc(c2)C)c1)c3cccc4cccc43</chem>	7.17	7.41
fl_269	<chem>O=C(c1cn(c2cccc21)CC)c3cccc4cccc43</chem>	5.87	5.53

Nome	Nomenclatura	<i>pKi</i> Exp_CB1	<i>pKi</i> Exp_CB2
fl_270	<chem>FC(F)(F)c1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)c1</chem>	6.61	7.15
fl_271	<chem>Fc1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)c1</chem>	7.8	08.04
fl_272	<chem>Clc1cccc1-c2cc(cn2CCCC)C(=O)c3cccc4cccc43</chem>	8.1	8.28
fl_273	<chem>O=C(c1cn(CCCCC)c(-c2ccc(CCCC)cc2)c1)c3cccc4cccc43</chem>	7.38	7.19
fl_275	<chem>O=C(c1cn(c(-c2cccc2CCCC)c1)CCCC)c3cccc4cccc43</chem>	7.22	7.16
fl_276	<chem>O=C(c1c2cccc2n(CCCCC)c1)Cc3cccc3</chem>	07.05	6.80
fl_277	<chem>O=C(c1c2cccc2n(CCCCC)c1C)Cc3cccc3</chem>	6.91	6.74
fl_278	<chem>BrC1ccc(CC(=O)c2c3cccc3n(CCCCC)c2)cc1</chem>	5.99	6.18
fl_279	<chem>BrC1cccc1CC(=O)c2c3cccc3n(CCCCC)c2</chem>	08.08	7.70
fl_280	<chem>O=C(c1cn(CCCC)c2cccc21)c3cccc4cccc43</chem>	08.05	7.42
fl_281	<chem>O=C(Cc1cccc1OC)c2c3cccc3n(CCCCC)c2</chem>	7.96	7.48
fl_282	<chem>O=C(c1cn(CCCCC)c2cccc21)Cc3cccc3C</chem>	7.54	6.84
fl_283	<chem>O=C(c1c2cccc2n(CCCCC)c1)Cc3cccc(OC)c3</chem>	7.77	07.05
fl_284	<chem>Clc1cccc(CC(=O)c2c3cccc3n(CCCCC)c2C)c1</chem>	6.93	6.86
fl_285	<chem>BrC1cccc1CC(=O)c2c3cccc3n(CCCCC)c2C</chem>	7.82	7.54
fl_286	<chem>Fc1ccc(CC(=O)c2c3cccc3n(CCCCC)c2)cc1</chem>	6.37	6.44
fl_287	<chem>Fc1ccc(CC(=O)c2c3cccc3n(CCCCC)c2C)cc1</chem>	5.54	6.11
fl_288	<chem>Oc1cc(CCCCC)cc2c1-c3cc(ccc3C(O2)(C)C)C</chem>	6.51	07.02
fl_289	<chem>FCCCCc1cc(O)c2c(OC([C@@H]3CC=C(C[C@@H]23)C)(C)C)c1</chem>	7.24	08.06
fl_290	<chem>O[C@@H]1CC[C@H]([C@@H](C1)c2ccc(C(CCCCC)(C)C)cc2O)CCCO</chem>	9.23	8.74

Anexo B - Estruturas e valores de pK_i que atuam no CB1R

Nome	Smiles	pK_i Exp_cb1
cb1_1	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccccc43</chem>	7.32
cb1_2	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3cccc4ccccc43</chem>	6.47
cb1_3	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3cccc4ccccc43</chem>	08.05
cb1_4	<chem>O=C(c1c(n(c2ccccc21)C)C)c3cccc4ccccc43</chem>	5
cb1_5	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.23
cb1_6	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.97
cb1_7	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.26
cb1_8	<chem>O=C(c1cn(c2ccccc21)CC)c3cccc4ccccc43</chem>	5.87
cb1_9	<chem>O=C(c1cn(CCCC)c2ccccc21)c3cccc4ccccc43</chem>	08.05
cb1_10	<chem>O=C(c1c2ccccc2n(CCC)c1)c3cccc4ccc(cc34)C</chem>	6.67
cb1_11	<chem>O=C(c1cn(c2ccccc21)CC)c3ccc(OC)c4ccccc43</chem>	06.09
cb1_12	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(OC)c4ccccc43</chem>	8.92
cb1_13	<chem>O=C(c1c2ccccc2n(c1CCCC)CCC)c3ccc(OC)c4ccccc43</chem>	7.39
cb1_14	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(OC)c4ccccc43</chem>	6.32
cb1_15	<chem>O=C(c1c2ccccc2n(c1CCCC)CCCC)c3ccc(OC)c4ccccc43</chem>	6.85
cb1_16	<chem>O=C(c1c2ccccc2n(c1CCCCC)CCCC)c3ccc(OC)c4ccccc43</chem>	6.34
cb1_17	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(OC)c4ccccc43</chem>	7.45
cb1_18	<chem>O=C(c1c2ccccc2n(c1CC)CCCC)c3cccc4ccccc43</chem>	7.28
cb1_19	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(c4ccccc43)C</chem>	9.16
cb1_20	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(c4ccccc43)C</chem>	8.3
cb1_21	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4ccc(OC)cc43</chem>	8.18
cb1_22	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4cc(OC)ccc34</chem>	7.36
cb1_23	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(c4ccccc43)CCC</chem>	7.59
cb1_24	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc34)CCC</chem>	7.28
cb1_25	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc34)CC</chem>	7.15
cb1_26	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(c4ccccc34)CC</chem>	7.48
cb1_27	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(c4ccccc34)CC</chem>	8.82
cb1_28	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4ccc(CC)cc34</chem>	08.08
cb1_29	<chem>O=C(c1c2ccccc2n(CCC)c1)c3cccc4ccc(CC)cc34</chem>	6.47
cb1_30	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4ccc(CC)cc43</chem>	5.87

Nome	Smiles	<i>pKi</i> Exp_cb1
cb1_31	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(CCCC)c4ccccc34</chem>	6.83
cb1_32	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(OCC)c4ccccc43</chem>	8.34
cb1_33	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccc(CC)cc43</chem>	7.55
cb1_34	<chem>Brc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	6.43
cb1_35	<chem>Clc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	6.73
cb1_36	<chem>Clc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	07.03
cb1_37	<chem>c1ccc(c2ccccc12)C(=O)c3cn(CCCCC)c4ccccc43</chem>	8.14
cb1_38	<chem>Fc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	6.62
cb1_39	<chem>Fc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	6.28
cb1_40	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc43)C</chem>	8.8
cb1_41	<chem>F[C@@H](CCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc43)C)C</chem>	8.49
cb1_42	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc34)CC</chem>	9.42
cb1_43	<chem>FCCCCn1c2ccccc2c(n1)C(=O)c3cccc4ccccc43</chem>	8.87
cb1_44	<chem>FCCCCn1cc(c2ccccc21)C(Oc3cccc4ccccc43)=O</chem>	9.33
cb1_45	<chem>Fc1ccccc1-n2c3ccccc3c(n2)C(Oc4ccc5ccccc5c4)=O</chem>	5
cb1_46	<chem>FCCCCn1c2ccccc2c(n1)C(Oc3cccc4ccccc34)=O</chem>	8.59
cb1_47	<chem>FCCCCn1cc(c2ccccc21)C(=O)NC34CC5CC(C3)CC(C4)C5</chem>	8.6
cb1_48	<chem>O=C(NCc1ccccc1)c2c3ccccc3n(CCCCC)c2</chem>	7.28
cb1_49	<chem>FCCCCn1cc(c2ccccc21)C(=O)NCc3ccccc3</chem>	7.14
cb1_50	<chem>O=C(Nc1ccccc1)c2cn(CCCCC)c3ccccc32</chem>	6.79
cb1_51	<chem>O=C(Nc1cccc2ccccc21)c3cn(CCCCC)c4ccccc43</chem>	8.74
cb1_52	<chem>FCCCCn1cc(c2ccccc21)C(=O)Nc3cccc4ccccc43</chem>	8.43
cb1_53	<chem>ClCCCCn1cc(c2ccccc21)C(=O)Nc3cccc4ccccc43</chem>	8.43
cb1_54	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)NC45CC6CC(C4)CC(C5)C6)cc1</chem>	8.97
cb1_55	<chem>O=C(Nc1cccc2ccccc21)c3c4ccccc4n(n3)CCCC</chem>	8.41
cb1_56	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3ccccc3OC</chem>	7.58
cb1_57	<chem>O=C(c1c2ccccc2n([C@H](CCCC)C)c1C)c3cccc4ccccc43</chem>	7.32
cb1_58	<chem>Brc1cccc2cccc(c12)C(=O)c3cn(CCCCC)c4ccccc43</chem>	7.68
cb1_59	<chem>O=C(c1cn(CCCCC#N)c2ccccc21)c3cccc4ccccc43</chem>	9.55
cb1_60	<chem>O=C(c1ccc(OCCCC)c2ccccc12)c3cccc4ccccc43</chem>	7.82
cb1_61	<chem>FCCCCn1cc(c2ccccc21)C(=O)NC(c3ccccc3)(C)C</chem>	7.3

Nome	Smiles	<i>pKi</i> Exp_cb1
cb1_62	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC)C(C)C</chem>	8.54
cb1_63	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)C)C(=O)N)cc1</chem>	09.05
cb1_64	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)(C)C)C(OC)=O)cc1</chem>	8.94
cb1_65	<chem>O=C(C1C(C1(C)C)(C)C)c2c3ccccc3n(CC4CCOCC4)c2</chem>	7.6
cb1_66	<chem>O=C(C12CC3CC(C1)CC(C2)C3)c4c5ccccc5n(CCCCC)c4</chem>	7.48
cb1_67	<chem>O=C(C1C(C1(C)C)(C)C)c2c3ccccc3n(CCCCC)c2</chem>	7.25
cb1_68	<chem>FCCCCn1c(c(C(=O)C2C(C2(C)C)(C)C)c3ccccc31)C</chem>	7.71
cb1_69	<chem>FCCCCn1c2ccccc2c(n1)C(=O)C3C(C3(C)C)(C)C</chem>	7.76
cb1_70	<chem>FC(F)(F)CCn1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31</chem>	7.36
cb1_71	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2)c1)c3cccc4cccc43</chem>	7.68
cb1_72	<chem>O=C(c1cn(CCCC)c(-c2ccccc2)c1)c3cccc4cccc43</chem>	7.22
cb1_73	<chem>O=C(c1cn(CCCCC)c(-c2ccc(OC)cc2)c1)c3cccc4cccc43</chem>	6.55
cb1_74	<chem>O=C(c1cn(CCCCC)c(-c2ccc(cc2)C)c1)c3cccc4cccc43</chem>	6.89
cb1_75	<chem>Clc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)cc1</chem>	6.56
cb1_76	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2OC)c1)c3cccc4cccc43</chem>	7.54
cb1_77	<chem>Fc1cccc1-c2cc(cn2CCCC)C(=O)c3cccc4cccc43</chem>	8.11
cb1_78	<chem>O=C(c1cc(n(CCCCC)c1)-c2cccc3ccccc32)c4cccc5ccccc54</chem>	7.39
cb1_79	<chem>O=C(c1cn(CCCCC)c(-c2ccc(c2)C)c1)c3cccc4cccc43</chem>	7.17
cb1_80	<chem>FC(F)(F)c1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)c1</chem>	6.61
cb1_81	<chem>Fc1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)c1</chem>	7.8
cb1_82	<chem>Clc1cccc1-c2cc(cn2CCCC)C(=O)c3cccc4cccc43</chem>	8.1
cb1_83	<chem>O=C(c1cn(CCCCC)c(-c2ccc(CCCC)cc2)c1)c3cccc4cccc43</chem>	7.38
cb1_84	<chem>FC(F)(F)c1cccc1-c2cc(cn2CCCC)C(=O)c3cccc4cccc43</chem>	7.11
cb1_85	<chem>O=C(c1cn(c(-c2ccccc2CCCC)c1)CCCC)c3cccc4cccc43</chem>	7.22
cb1_86	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccccc3</chem>	07.05
cb1_87	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccccc3</chem>	6.91
cb1_88	<chem>Brc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2)cc1</chem>	5.99
cb1_89	<chem>Brc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	08.08
cb1_90	<chem>O=C(Cc1cccc1OC)c2c3ccccc3n(CCCCC)c2</chem>	7.96
cb1_91	<chem>O=C(c1cn(CCCCC)c2ccccc21)Cc3ccccc3C</chem>	7.54
cb1_92	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3cccc(OC)c3</chem>	7.77

Nome	Smiles	<i>pKi</i> Exp_cb1
cb1_124	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)(C)C</chem>	9.54
cb1_125	<chem>O=C(NC1CC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CCCC#N)C(C)(C)C</chem>	8.54
cb1_126	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)N)c1</chem>	8.87
cb1_127	<chem>O=C(N[C@@H](CC(C)C)C(=O)N)c1c2ccccc2n(n1)CC3CCCC3</chem>	8.47
cb1_128	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)C)C(=O)N)c1</chem>	8.37
cb1_129	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCOCC3)C(C)C</chem>	7.38
cb1_130	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NC4CC4)c1</chem>	7.58
cb1_131	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C=4[N-N=NN4)c1</chem>	7.89
cb1_132	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NC[C@@H](O)CO)c1</chem>	7.52
cb1_133	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCOCC3)C(C)(C)C</chem>	8.73
cb1_134	<chem>O=C(N[C@@H](CC(C)C)C(=O)N)c1c2ccccc2n(n1)CC3CCOCC3</chem>	7.34
cb1_135	<chem>Fc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)c1</chem>	9.76
cb1_136	<chem>O=C(OCC)[C@@H]1CCCC[C@@H]1NC(=O)c2c3ccccc3n(n2)CC4CCOCC4</chem>	7.86
cb1_137	<chem>Fc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)c1</chem>	9.84
cb1_138	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)(C)C</chem>	8.35
cb1_139	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)C</chem>	7.96
cb1_140	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)(C)C</chem>	9.24
cb1_141	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NC3CC3)CC4CCOCC4</chem>	8.21
cb1_142	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)CC3CCOCC3</chem>	9.15
cb1_143	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)C[C@@H]3CCCCO3)C(C)(C)C</chem>	8.39
cb1_144	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)CC3CCOCC3</chem>	8.76
cb1_145	<chem>Fc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@@H](C(=O)NCCO)C(C)(C)C)c1</chem>	9.46
cb1_146	<chem>Fc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)c1</chem>	8.21
cb1_147	<chem>OC[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)c4ccccc4</chem>	6.93
cb1_148	<chem>Fc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@@H](C(=O)NC4CC4)C(C)(C)C)c1</chem>	8.4
cb1_149	<chem>Fc1ccc2c(c(nn2CCCC#N)C(=O)N[C@@H](C(C)(C)C)C(=O)N)c1</chem>	08.05
cb1_150	<chem>O=C(N[C@@H]1CCCC[C@@H]1OCc2ccccc2)c3c4ccccc4n(n3)CC5CCCC5</chem>	07.05
cb1_151	<chem>Fc1ccc2c(c(nn2CCCC#N)C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)c1</chem>	8.22
cb1_152	<chem>O=C(NC1CC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CC4CCCC4)C(C)C</chem>	8.25
cb1_153	<chem>O=C(NCC([O-])=O)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)(C)C</chem>	8.69
cb1_154	<chem>FC(F)(F)Oc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)c1</chem>	7.48

Nome	Smiles	<i>pKi</i> Exp_cb1
cb1_155	<chem>O=C(NCC(=O)N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)C</chem>	7.68
cb1_156	<chem>FC(F)(F)Oc1ccc2c(c(nn2CC3CCCCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)c1</chem>	9.32
cb1_157	<chem>O=C(NCCO)[C@@H](NC(=O)c1c2cc(OC)ccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.12
cb1_158	<chem>O=C(N)[C@@H](NC(=O)c1c2cc(OC)ccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	8.97
cb1_159	<chem>O[C@@H](CNC(=O)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCOCC3)C(C)(C)C)CO</chem>	7.83
cb1_160	<chem>FC(F)(F)c1c(CNC(=O)c2c3ccccc3n(n2)CC4CCOCC4)cc(o1)C</chem>	8.76
cb1_161	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCC3CCCCC3)CC4CCCCC4</chem>	7.63
cb1_162	<chem>Fc1cccc2c(nn(CC3CCCCC3)c12)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCc4ccccc4</chem>	8.4
cb1_163	<chem>Clc1ccc2c(c(nn2CCCC#N)C(=O)N[C@@H](C(C)(C)C)C(=O)N)c1</chem>	8.25
cb1_164	<chem>Clc1ccc2c(c(nn2CCCC#N)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)c1</chem>	7.79
cb1_165	<chem>Clc1ccc2c(c(nn2CCCC#N)C(=O)N[C@@H](C(C)(C)C)C(OC)=O)c1</chem>	8.36
cb1_166	<chem>O=C(N[C@@H](C(C)(C)C)C(OC)=O)c1c2cccnc2n(n1)CC3CCOCC3</chem>	8.29
cb1_167	<chem>Fc1cccc2c(nn(CC3CCCCC3)c12)C(=O)N[C@H](C(=O)NCc4ccc(cc4)C(OC)=O)C(C)(C)C</chem>	7.78
cb1_168	<chem>O=C(OC)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCOCC3)C(C)(C)C</chem>	9.23
cb1_169	<chem>Fc1cccc2c(nn(CC3CCCCC3)c12)C(=O)N[C@H](C(=O)NCc4ccc(OC)c4)C(C)(C)C</chem>	8.44
cb1_170	<chem>O=C(NC1CC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CC4CCCCC4)C(C)(C)C</chem>	9.55
cb1_171	<chem>O=C(NC)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.71
cb1_172	<chem>O=C(NCCC([O-])=O)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)(C)C</chem>	7.72
cb1_173	<chem>O=C(N[C@@H](C(C)(C)C)c1nnc(o1)N)c2c3ccccc3n(n2)CCCC#N</chem>	9.89
cb1_174	<chem>O=C(NCC(=O)N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	8.94
cb1_175	<chem>O=C(NCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.73
cb1_176	<chem>Fc1ccc2c(nn(CC3CCCCC3)c2c1)C(=O)N[C@H](C(=O)NC4CC4)C(C)(C)C</chem>	9.32
cb1_177	<chem>O=C(NC1CC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CC4CCOCC4)C(C)(C)C</chem>	8.63
cb1_178	<chem>O=C(NC1CC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CCCC#C)C(C)(C)C</chem>	08.02
cb1_179	<chem>O=C(NCC(=O)N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCOCC3)C(C)(C)C</chem>	7.91
cb1_180	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCCCO)CC3CCOCC3)c1</chem>	09.04
cb1_181	<chem>O=C(NC1CCC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CCCC#C)C(C)(C)C</chem>	7.84
cb1_182	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)CC3CCOCC3)c1</chem>	8.17
cb1_183	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)N)CC3CCOCC3)c1</chem>	8.32
cb1_184	<chem>O=C(NCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)(C)C</chem>	08.02
cb1_185	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)CCCC#N)c1</chem>	9.19

Nome	Smiles	<i>pKi</i> Exp_cb1
cb1_186	<chem>Clc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2C)cc1</chem>	5.8
cb1_187	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4ccc(cc43)C</chem>	6.46
cb1_188	<chem>O=C(N[C@@H](Cc1ccccc1)C(=O)N)c2c3ccccc3n(n2)CCCC#N</chem>	5.61
cb1_189	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc43)C</chem>	6.91
cb1_190	<chem>Fc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	7.39
cb1_191	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccc(OC)cc3</chem>	5.97
cb1_192	<chem>Fc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2)c1</chem>	7.14
cb1_193	<chem>Clc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2C</chem>	7.89
cb1_194	<chem>O=C(N[C@@H]1c2ccccc2CCC1)c3c4ccnnc4n(n3)CC5CCOCC5</chem>	7.67
cb1_195	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(OC)c4ccccc43</chem>	7.2
cb1_196	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccc(C)cc3</chem>	6.75
cb1_197	<chem>Clc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2)c1</chem>	7.42
cb1_198	<chem>O=C(Nc1c(C)c(no1)C)c2c3ccccc3n(n2)CC4CCOCC4</chem>	7.34
cb1_199	<chem>Clc1ccc(C(=O)c2c3ccccc3n(CCCCC)c2C)c4ccccc14</chem>	08.05
cb1_200	<chem>FCCCCN1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31</chem>	7.53
cb1_201	<chem>Fc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	7.64
cb1_202	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3ccc(OC)c4ccccc43</chem>	7.47
cb1_203	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccccc3OC</chem>	7.6
cb1_204	<chem>O=C(Oc1cccc2ccccc12)c3c4ccccc4n(n3)CCCC</chem>	8.44
cb1_205	<chem>O=C(c1cc(n(CCCCC)c1)-c2ccc3ccccc3c2)c4cccc5ccccc54</chem>	6.48
cb1_206	<chem>Clc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	8.1
cb1_207	<chem>FCCCCN1cc(c2ccccc21)C(=O)Nc3ccc4ccccc4c3</chem>	6.63
cb1_208	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC)C(C)(C)C</chem>	9.23
cb1_209	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2)c1)c3cccc4ccccc43</chem>	7.85
cb1_210	<chem>O=C(c1cn(CCCCC)c(-c2ccc(OC)c2)c1)c3cccc4ccccc43</chem>	7.28
cb1_211	<chem>O=C(c1cn(CCCC)c2ccccc21)c3ccc(OC)c4ccccc43</chem>	8.25
cb1_212	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(c4ccccc43)CCC</chem>	9.19
cb1_213	<chem>O=C(NCCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)C[C@H]3CCC(O3)=O)C(C)(C)C</chem>	8.84
cb1_214	<chem>O=C(N[C@@H](C(C)(C)C)C(=O)N)c1c2cc(ccc2n(n1)CCCC#N)C</chem>	8.24
cb1_215	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3cccc4ccccc43</chem>	08.01
cb1_216	<chem>FCCCCN1cc(c2ccccc21)C(=O)c3cccc4ccccc43</chem>	9

Nome	Smiles	pKi Exp_cb1
cb1_217	<chem>O=C(NCCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)C</chem>	7.82
cb1_218	<chem>O=C(NC1CCC1)[C@@H](NC(=O)c2c3ccccc3n(n2)CC4CCOCC4)C(C)(C)C</chem>	8.18
cb1_219	<chem>O=C(NCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCOCC3)C(C)(C)C</chem>	8.64
cb1_220	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCCCO)CCCC#N)c1</chem>	8.7
cb1_221	<chem>O=C(NC1CC1)[C@@H](NC(=O)c2c3cc(OC)ccc3n(n2)CC4CCCC4)C(C)(C)C</chem>	8.35
cb1_222	<chem>Fc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)N)c1</chem>	8.79
cb1_223	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NC3CC3)CCCC#C</chem>	8.55
cb1_224	<chem>O=C(NCCC([O-])=O)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)(C)C</chem>	9.13
cb1_225	<chem>Fc1cccc2c(nn(CC3CCCC3)c12)C(=O)N[C@H](C(=O)NCc4ccc(cc4)C#N)C(C)(C)C</chem>	7.83
cb1_226	<chem>Brc1ccc(c2ccccc12)C(=O)c3cn(CCCC)c4cccc43</chem>	8.92
cb1_227	<chem>O=C(NCCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)(C)C</chem>	9.27
cb1_228	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)c1</chem>	8.6
cb1_229	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC(=O)C(C)(C)C(C)(C)C</chem>	7.98
cb1_230	<chem>Clc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCC(=O)N)c1</chem>	7.76
cb1_231	<chem>Fc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)c1</chem>	8.88
cb1_232	<chem>Fc1ccc2c(c(nn2CC3CCOCC3)C(=O)N[C@@H](C(C)(C)C)C(=O)NC4CC4)c1</chem>	9.47
cb1_233	<chem>Fc1cccc2c(nn(CC3CCCC3)c12)C(=O)N[C@H](C(=O)NCc4ccc(cc4)C([O-])=O)C(C)(C)C</chem>	8.39
cb1_234	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3cccc4cccc43</chem>	08.02
cb1_235	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4cccc43</chem>	7.66
cb1_236	<chem>O=C(c1c2ccccc2n(c1C)CC)c3cccc4cccc43</chem>	5.93
cb1_237	<chem>O=C(c1cn(c2ccccc21)C)c3cccc4cccc43</chem>	5.00
cb1_238	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3ccc(OC)c4cccc43</chem>	8.35
cb1_239	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(CCC)c4cccc34</chem>	6.47
cb1_240	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3ccc(CCC)c4cccc34</chem>	7.38
cb1_241	<chem>Brc1ccc(c2ccccc12)C(=O)c3cn(c4cccc43)CCC</chem>	6.79
cb1_242	<chem>Fc1ccc(C(=O)c2c3ccccc3n(CCCC)c2C)c4cccc14</chem>	7.85
cb1_243	<chem>O=C(c1c2ccccc2n(n1)CCCC)c3cccc4cccc43</chem>	8.23
cb1_244	<chem>Fc1ccc(Cn2cc(c3ccccc32)C(Oc4cccc5ccccc45)=O)cc1</chem>	8.92
cb1_245	<chem>O=C(NC12CC3CC(C1)CC(C2)C3)c4c5ccccc5n(CCCC)c4</chem>	8.19
cb1_246	<chem>FCCCCN1c2ccccc2c(n1)C(=O)NC34CC5CC(C3)CC(C4)C5</chem>	8.71
cb1_247	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(CCCC)c1)C(C)(C)C</chem>	8.89

Nome	Smiles	p <i>K</i> _i Exp_cb1
cb1_248	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)C</chem>	9.11
cb1_249	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)C)C(OC)=O)cc1</chem>	8.00
cb1_250	<chem>F[C@@H](Cn1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31)CCC</chem>	7.23
cb1_251	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2)c1)c3cccc4ccccc43</chem>	7.96
cb1_252	<chem>Clc1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)c1</chem>	7.15
cb1_253	<chem>FC(F)(F)c1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	6.66
cb1_254	<chem>O=C(c1cn(CCCCC)c(-c2ccc(CC)cc2)c1)c3cccc4ccccc43</chem>	7.47
cb1_255	<chem>O=C(c1cn(c(-c2ccccc2CC)c1)CCCC)c3cccc4ccccc43</chem>	7.77
cb1_256	<chem>O=C(c1cn(c(-c2ccccc2C)c1)CCCC)c3cccc4ccccc43</chem>	8.25
cb1_257	<chem>Clc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2)cc1</chem>	6.41
cb1_258	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccc(C)cc3</chem>	6.13
cb1_259	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3cccc(OC)c3</chem>	7.21
cb1_260	<chem>Fc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2C</chem>	7.41
cb1_261	<chem>O=C(OC)[C@H](NC(=O)c1c2ccccc2n(CCCCC)c1)C(C)C</chem>	7.82
cb1_262	<chem>ClCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)C)C(=O)N</chem>	8.39
cb1_263	<chem>FCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)(C)C)C(=O)N</chem>	8.84
cb1_264	<chem>FCCCCn1c2ccccc2c(n1)C(=O)N[C@H](Cc3ccccc3)C(=O)N</chem>	6.90
cb1_265	<chem>O=C(OCC)[C@@H]1CCCC[C@@H]1NC(=O)c2c3ccccc3n(n2)CC4CCCC4</chem>	8.42
cb1_266	<chem>O=C(N[C@@H](Cc1cccc1)C(=O)N)c2c3ccccc3n(n2)CC4CCOCC4</chem>	6.56
cb1_267	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCCCO)CCCC#N</chem>	7.90
cb1_268	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)N)CCCC#N</chem>	8.59
cb1_269	<chem>OC[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)(C)C</chem>	8.51
cb1_270	<chem>FC(F)(F)Oc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@H](C(=O)NCCCO)C(C)(C)C)c1</chem>	08.06
cb1_271	<chem>O=C(N[C@@H](C(C)(C)C)C(=O)N)c1c2cccnc2n(n1)CCCC#N</chem>	7.34
cb1_272	<chem>FC(F)(F)Oc1ccc2c(c(nn2CC3CCCC3)C(=O)N[C@H](C(=O)NC4CC4)C(C)(C)C)c1</chem>	9.44
cb1_273	<chem>O=C(NCCCO)[C@@H](NC(=O)c1c2cc(OC)ccc2n(n1)CC3CCCC3)C(C)(C)C</chem>	8.77
cb1_274	<chem>O=C(N[C@@H](C(C)(C)C)C(=O)NC1CC1)c2c3cc(c3n(n2)CCCC#N)C</chem>	7.60
cb1_275	<chem>Fc1cccc2c(nn(CCC3CCCC3)c12)C(=O)N[C@H](C(=O)NCCc4ccccc4)C(C)(C)C</chem>	8.61
cb1_276	<chem>Fc1cccc2c(nn(CCC3CCCC3)c12)C(=O)N[C@H](C(=O)NCc4cnccc4)C(C)(C)C</chem>	7.53
cb1_277	<chem>O=C(OC)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#N)C(C)(C)C</chem>	9.15
cb1_278	<chem>O=C(OC)[C@@H](NC(=O)c1c2ccccc2n(n1)CC(OC(C)(C)C)=O)C(C)(C)C</chem>	8.99

Nome	Smiles	<i>pKi</i> Exp_cb1
cb1_279	<chem>O=C(NC(C)C)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.34
cb1_280	<chem>Fc1ccc2c(nn(CC3CCCCC3)c2c1)C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO</chem>	9.27
cb1_281	<chem>O=C(NCCOC)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.21
cb1_282	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NCCO)CC3CCOCC3)c1</chem>	7.81
cb1_283	<chem>O=C(NCCO)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC#C)C(C)(C)C</chem>	8.45
cb1_284	<chem>FC(F)(F)CCCCc1cc(O)c2c(OC([C@@H]3CC=C(C[C@@H]23)C)(C)C)c1</chem>	7.70
cb1_285	<chem>O=C(c1cn(CCCCC)cc1)c2cccc3cccc32</chem>	07.06
cb1_286	<chem>OC[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)C</chem>	7.31
cb1_287	<chem>Fc1cccc2c1n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)N)CC3CCOCC3</chem>	7.56
cb1_288	<chem>Fc1ccc2c(nn(CC3CCCCC3)c2c1)C(=O)N[C@@H](C(C)(C)C)C(=O)N</chem>	8.32
cb1_289	<chem>Fc1ccc2c(n(nc2C(=O)N[C@@H](C(C)(C)C)C(=O)NC3CC3)CC4CCOCC4)c1</chem>	8.15
cb1_290	<chem>O=C(Oc1cccc2cccnc21)c3c4ccccc4n(CC5CCCCC5)c3</chem>	9.66

Anexo C - Estruturas e valores de pK_i que atuam no CB2R

Nome	Smiles	pK_i Exp_cb2
cb2_1	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccccc43</chem>	8.40
cb2_2	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3cccc4ccccc43</chem>	8.53
cb2_3	<chem>O=C(c1c(n(c2ccccc21)C)C)c3cccc4ccccc43</chem>	5.30
cb2_4	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3cccc4ccc(cc43)C</chem>	8,46
cb2_5	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccc(cc43)C</chem>	7.49
cb2_6	<chem>O=C(c1cn(CCCC)c2ccccc21)c3cccc4ccccc43</chem>	7.42
cb2_7	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(OC)c4ccccc43</chem>	7.91
cb2_8	<chem>O=C(c1c2ccccc2n(c1CCCC)CCC)c3ccc(OC)c4ccccc43</chem>	7.23
cb2_9	<chem>O=C(c1c2ccccc2n(c1CCCCC)CCCC)c3ccc(OC)c4ccccc43</chem>	6.92
cb2_10	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4ccc(OC)cc43</chem>	8.16
cb2_11	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(c4ccccc43)CCC</chem>	08.02
cb2_12	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc4)CCC</chem>	7.92
cb2_13	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc4)CC</chem>	7.92
cb2_14	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(c4ccccc4)CC</chem>	8.00
cb2_15	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(c4ccccc4)CC</chem>	9.38
cb2_16	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4ccc(CC)cc43</chem>	6.62
cb2_17	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(CCCC)c4ccccc43</chem>	7.31
cb2_18	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccc(CC)cc43</chem>	8.25
cb2_19	<chem>Clc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	7.66
cb2_20	<chem>Clc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	7.36
cb2_21	<chem>Fc1ccc(c2ccccc12)C(=O)c3cn(CCCCC)c4ccccc43</chem>	8.49
cb2_22	<chem>Fc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	7.48
cb2_23	<chem>Fc1ccc(C(=O)c2c3ccccc3n(c2C)CCC)c4ccccc14</chem>	7.42
cb2_24	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc43)C</chem>	9.24
cb2_25	<chem>F[C@@H](CCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc43)C)C</chem>	09.09
cb2_26	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3ccc(c4ccccc4)CC</chem>	9.43
cb2_27	<chem>FCCCCn1c2ccccc2c(n1)C(=O)c3cccc4ccccc43</chem>	8.88
cb2_28	<chem>FCCCCn1cc(c2ccccc21)C(Oc3cccc4ccnc43)=O</chem>	9.20
cb2_29	<chem>Fc1ccc(Cn2cc(c3ccccc32)C(Oc4cccc5ccnc54)=O)cc1</chem>	9.32
cb2_30	<chem>FCCCCn1c2ccccc2c(n1)C(Oc3cccc4ccccc4)=O</chem>	8.47

Nome	Smiles	<i>pKi</i> Exp_cb2
cb2_31	<chem>FCCCCCn1cc(c2ccccc21)C(=O)NC34CC5CC(C3)CC(C4)C5</chem>	9.10
cb2_32	<chem>O=C(NCc1ccccc1)c2c3ccccc3n(CCCCC)c2</chem>	6.73
cb2_33	<chem>FCCCCCn1cc(c2ccccc21)C(=O)NCc3ccccc3</chem>	6.37
cb2_34	<chem>O=C(Nc1cccc2ccccc21)c3cn(CCCCC)c4ccccc43</chem>	7.66
cb2_35	<chem>FCCCCCn1cc(c2ccccc21)C(=O)Nc3cccc4ccccc43</chem>	7.87
cb2_36	<chem>ClCCCCCn1cc(c2ccccc21)C(=O)Nc3cccc4ccccc43</chem>	7.87
cb2_37	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)NC45CC6CC(C4)CC(C5)C6)cc1</chem>	9.76
cb2_38	<chem>O=C(Nc1cccc2ccccc21)c3c4ccccc4n(n3)CCCC</chem>	8.46
cb2_39	<chem>O=C(c1c2ccccc2n([C@H](CCCC)C)c1C)c3cccc4ccccc43</chem>	8.40
cb2_40	<chem>Brc1cccc2cccc(c12)C(=O)c3cn(CCCCC)c4ccccc43</chem>	8.26
cb2_41	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)(C)C)C(OC)=O)cc1</chem>	9.92
cb2_42	<chem>O=C(C12CC3CC(C1)CC(C2)C3)c4c5ccccc5n(CCCCC)c4</chem>	8.76
cb2_43	<chem>FCCCCCn1c2ccccc2c(n1)C(=O)C3C(C3(C)C)(C)C</chem>	9.35
cb2_44	<chem>FC(F)(F)CCn1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31</chem>	09.08
cb2_45	<chem>O=C(c1cn(CCCC)c(-c2ccccc2)c1)c3cccc4ccccc43</chem>	7.82
cb2_46	<chem>O=C(c1cn(CCCCC)c(-c2ccc(cc2)C)c1)c3cccc4ccccc43</chem>	7.74
cb2_47	<chem>Clc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	7.60
cb2_48	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2OC)c1)c3cccc4ccccc43</chem>	7.70
cb2_49	<chem>Fc1ccccc1-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43</chem>	8.48
cb2_50	<chem>O=C(c1cn(CCCCC)c(-c2cccc(c2)C)c1)c3cccc4ccccc43</chem>	7.41
cb2_51	<chem>O=C(c1cn(CCCCC)c(-c2ccc(CCC)cc2)c1)c3cccc4ccccc43</chem>	7.19
cb2_52	<chem>O=C(c1cn(c(-c2ccccc2CCCC)c1)CCCC)c3cccc4ccccc43</chem>	7.16
cb2_53	<chem>Brc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2)cc1</chem>	6.18
cb2_54	<chem>Brc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	7.70
cb2_55	<chem>O=C(c1cn(CCCCC)c2ccccc21)Cc3ccccc3C</chem>	6.84
cb2_56	<chem>Clc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2C)c1</chem>	6.86
cb2_57	<chem>Brc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2C</chem>	7.54
cb2_58	<chem>Fc1ccc(CC(=O)c2c3ccccc3n(CCCCC)c2)cc1</chem>	6.44
cb2_59	<chem>FCCCCCc1cc(O)c2c(OC([C@@H]3CC=C(C[C@@H]23)C)(C)C)c1</chem>	08.06
cb2_60	<chem>O[C@@H]1CC[C@H]([C@@H](C1)c2ccc(C(CCCCC)(C)C)cc2O)CCCO</chem>	8.74
cb2_61	<chem>O[C@H]1CC[C@@H]([C@H](C1)c2ccc(C(CCCCC)(C)C)cc2O)CCCO</chem>	7.63

Nome	Smiles	pKi Exp_cb2
cb2_62	<chem>O=C(NCCO)CCC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\C\CCCC</chem>	6.43
cb2_63	<chem>Clc1ccc(c2ccccc12)C(=O)c3cn(CCCCF)c4ccccc43</chem>	8.93
cb2_64	<chem>O=C(OC)[C@@H](OC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	8.95
cb2_65	<chem>FCCCCn1cc(C(=O)N[C@@H](C(C)C)C(=O)N)c2ccccc21</chem>	07.05
cb2_66	<chem>O=C(OC)[C@@H](NC(=O)c1c2ccccc2n(CC3CCCCC3)c1)C(C)(C)C</chem>	9.46
cb2_67	<chem>FCCCCn1cc(c2ccccc21)C(=O)N[C@H](C(C)C)C(OC)=O</chem>	7.70
cb2_68	<chem>O=C(NC(c1ccccc1)(C)C)c2c3ccccc3n(n2)CC4CCOCC4</chem>	8.86
cb2_69	<chem>Fc1cccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)C)C(=O)N)c1</chem>	7.28
cb2_70	<chem>FCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)C)C(=O)N</chem>	8.42
cb2_71	<chem>FCCCCn1c2ccccc2c(n1)C(=O)N[C@H]([C@H](CC)C)C(=O)N</chem>	8.37
cb2_72	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.48
cb2_73	<chem>O=C(OC)[C@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)C</chem>	9.52
cb2_74	<chem>lc1ccccc1C(=O)c2c3ccccc3n(CCCC)c2</chem>	7.31
cb2_75	<chem>O=C(OC)[C@@H](NC(=O)c1ccc2c(c3ccccc3n2CC4CCCCC4)c1)C(C)(C)C</chem>	8.18
cb2_76	<chem>O=C(c1ccc2c(c3ccccc3n2CCCC)c1)c4ccccc5ccccc54</chem>	8.64
cb2_77	<chem>FCCCCn1c2ccc(cc2c3ccccc31)C(=O)c4ccccc5ccccc54</chem>	8.36
cb2_78	<chem>O=C(c1c2ccccc2n(CCC)c1C)c3cccc4ccc(cc43)C</chem>	7.79
cb2_79	<chem>Fc1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	7.48
cb2_80	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccc(OC)cc3</chem>	6.35
cb2_81	<chem>Fc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2)c1</chem>	07.04
cb2_82	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(OC)c4ccccc43</chem>	7.49
cb2_83	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccc(C)cc3</chem>	6.24
cb2_84	<chem>FCCCCn1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31</chem>	9.22
cb2_85	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccccc3OC</chem>	07.09
cb2_86	<chem>O=C(Oc1cccc2ccccc12)c3c4ccccc4n(n3)CCCC</chem>	8.58
cb2_87	<chem>O=C(c1cc(n(CCCCC)c1)-c2ccc3ccccc3c2)c4ccccc5ccccc54</chem>	6.77
cb2_88	<chem>Clc1ccccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	8.15
cb2_89	<chem>O=C(c1cn(CCCCC)c(-c2cccc(OC)c2)c1)c3cccc4ccccc43</chem>	7.64
cb2_90	<chem>O=C(c1cn(CCCC)c2ccccc21)c3ccc(OC)c4ccccc43</chem>	8.66
cb2_91	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3cccc4ccccc43</chem>	8.26
cb2_92	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3cccc4ccccc43</chem>	8.59

Nome	Smiles	<i>pKi</i> Exp_cb2
cb2_93	<chem>Brc1ccc(c2ccccc12)C(=O)c3cn(CCCCC)c4ccccc43</chem>	8.96
cb2_94	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3cccc4ccccc43</chem>	8.53
cb2_95	<chem>O=C(c1c2ccccc2n(c1C)CC)c3cccc4ccccc43</chem>	06.02
cb2_96	<chem>O=C(c1cn(c2ccccc21)C)c3cccc4ccccc43</chem>	5.00
cb2_97	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(OC)c4ccccc43</chem>	8.73
cb2_98	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)c3ccc(CCCC)c4ccccc34</chem>	8.19
cb2_99	<chem>Brc1ccc(c2ccccc12)C(=O)c3cn(c4ccccc43)CCC</chem>	7.57
cb2_100	<chem>Fc1ccc(C(=O)c2c3ccccc3n(CCCCC)c2C)c4ccccc14</chem>	8.66
cb2_101	<chem>O=C(c1c2ccccc2n(n1)CCCC)c3cccc4ccccc43</chem>	8.34
cb2_102	<chem>FCCCCn1c2ccccc2c(n1)C(=O)NC34CC5CC(C3)CC(C4)C5</chem>	9.58
cb2_103	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)C)C(OC)=O)cc1</chem>	9.10
cb2_104	<chem>F[C@@H](Cn1cc(C(=O)C2C(C2(C)C)(C)C)c3ccccc31)CCC</chem>	8.74
cb2_105	<chem>O=C(c1cn(CCCCC)c(-c2ccccc2)c1)c3cccc4ccccc43</chem>	8.15
cb2_106	<chem>Clc1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)c1</chem>	7.80
cb2_107	<chem>FC(F)(F)c1ccc(-c2cc(cn2CCCC)C(=O)c3cccc4ccccc43)cc1</chem>	7.28
cb2_108	<chem>O=C(c1cn(CCCCC)c(-c2ccc(CC)cc2)c1)c3cccc4ccccc43</chem>	7.54
cb2_109	<chem>O=C(c1cn(c(-c2ccccc2C)c1)CCCC)c3cccc4ccccc43</chem>	8.40
cb2_110	<chem>O=C(c1c2ccccc2n(CCCCC)c1C)Cc3ccc(C)cc3</chem>	5.87
cb2_111	<chem>Fc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2C</chem>	7.12
cb2_112	<chem>ClCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)C)C(=O)N</chem>	7.92
cb2_113	<chem>FCCCCn1c2ccccc2c(n1)C(=O)N[C@H](Cc3ccccc3)C(=O)N</chem>	7.76
cb2_114	<chem>O=C(OC)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCCC3)C(C)(C)C</chem>	9.65
cb2_115	<chem>O=C(Oc1cccc2ccnc21)c3c4ccccc4n(CC5CCCCC5)c3</chem>	9.47
cb2_116	<chem>O=C(C1C(C1(C)C)(C)C)c2c3ccccc3n(CC4CCOCC4)c2</chem>	9.25
cb2_117	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CCCC)C(C)C</chem>	09.06
cb2_118	<chem>FCCCCn1cc(c2ccccc21)C(=O)c3ccc(F)c4ccccc43</chem>	8.72
cb2_119	<chem>Clc1ccc(C(=O)c2c3ccccc3n(CCCCC)c2C)c4ccccc14</chem>	8.64
cb2_120	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3ccccc3OC</chem>	8.54
cb2_121	<chem>O=C(c1cn(c(-c2ccccc2CC)c1)CCCC)c3cccc4ccccc43</chem>	8.47
cb2_122	<chem>O=C(c1c2ccccc2n(CCCCC)c1)c3cccc4ccc(CC)cc34</chem>	8.42
cb2_123	<chem>O=C(c1c2ccccc2n(CCCC)c1C)c3cccc4ccccc43</chem>	8.37

Nome	Smiles	<i>pKi</i> Exp_cb2
cb2_124	<chem>Clc1cccc1-c2cc(cn2CCCC)C(=O)c3cccc4cccc43</chem>	8.28
cb2_125	<chem>O=C(c1cn(CCCCC)c(-c2cccc2)c1)c3cccc4cccc43</chem>	8.19
cb2_126	<chem>O=C(c1c2cccc2n(CCCC)c1C)c3ccc(OC)c4cccc43</chem>	7.88
cb2_127	<chem>O=C(OC)[C@H](NC(=O)c1c2cccc2n(CCCCC)c1)C(C)C</chem>	7.85
cb2_128	<chem>O=C(c1c2cccc2n(CCCCC)c1C)c3ccc(OC)c4cccc43</chem>	7.75
cb2_129	<chem>Clc1cccc1CC(=O)c2c3cccc3n(CCCCC)c2C</chem>	7.60
cb2_130	<chem>Brc1ccc(C(=O)c2c3cccc3n(c2C)CCC)c4cccc14</chem>	7.51
cb2_131	<chem>O=C(Cc1cccc1OC)c2c3cccc3n(CCCCC)c2</chem>	7.48
cb2_132	<chem>O=C(c1cn(CCCCC)c(-c2ccc(OC)cc2)c1)c3cccc4cccc43</chem>	7.39
cb2_133	<chem>O=C(c1cc(n(CCCCC)c1)-c2cccc3cccc32)c4cccc5cccc54</chem>	7.31
cb2_134	<chem>O=C(c1cn(CCCCC)c(-c2cccc2)c1)c3cccc4cccc43</chem>	7.21
cb2_135	<chem>FC(F)(F)c1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)c1</chem>	7.15
cb2_136	<chem>O=C(c1c2cccc2n(CCCCC)c1C)Cc3cccc(OC)c3</chem>	07.08
cb2_137	<chem>O=C(c1c2cccc2n(c1C)CCC)c3ccc(OC)c4cccc43</chem>	07.01
cb2_138	<chem>O=C(c1c2cccc2n(CCCCC)c1C)Cc3cccc3</chem>	6.74
cb2_139	<chem>O=C(c1c2cccc2n(c1CCCC)CCC)c3ccc(OC)c4cccc43</chem>	6.51
cb2_140	<chem>Clc1ccc(CC(=O)c2c3cccc3n(CCCCC)c2)cc1</chem>	6.30
cb2_141	<chem>Fc1ccc(CC(=O)c2c3cccc3n(CCCCC)c2C)cc1</chem>	6.11
cb2_142	<chem>Clc1ccc(CC(=O)c2c3cccc3n(CCCCC)c2C)cc1</chem>	5.43
cb2_143	<chem>O=C(c1c2cccc2n(c1C)CCC)c3cccc4cccc43</chem>	7.86
cb2_144	<chem>O=C(c1c2cccc2n(CCCCC)c1C)c3cccc4ccc(cc43)C</chem>	9.31
cb2_145	<chem>O=C(c1cn(c2cccc21)CC)c3cccc4cccc43</chem>	5.53
cb2_146	<chem>O=C(c1c2cccc2n(CCC)c1)c3cccc4ccc(cc34)C</chem>	6.97
cb2_147	<chem>O=C(c1cn(c2cccc21)CC)c3ccc(OC)c4cccc43</chem>	6.20
cb2_148	<chem>O=C(c1c2cccc2n(CCC)c1)c3cccc4ccc(CC)cc34</chem>	6.91
cb2_149	<chem>Fc1cccc1-n2c3cccc3c(n2)C(Oc4ccc5cccc5c4)=O</chem>	6.37
cb2_150	<chem>O=C(Nc1cccc1)c2cn(CCCCC)c3cccc32</chem>	6.56
cb2_151	<chem>FCCCCn1cc(c2cccc21)C(=O)NC(c3cccc3)(C)C</chem>	7.10
cb2_152	<chem>O=C(C1C(C1(C)C)(C)C)c2c3cccc3n(CCCCC)c2</chem>	8.83
cb2_153	<chem>FCCCCn1c(c(C(=O)C2C(C2(C)C)(C)C)c3cccc31)C</chem>	8.34
cb2_154	<chem>Fc1cccc(-c2cc(cn2CCCC)C(=O)c3cccc4cccc43)c1</chem>	08.04

Nome	Smiles	<i>pKi</i> Exp_cb2
cb2_155	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3ccccc3</chem>	6.80
cb2_156	<chem>O=C(c1c2ccccc2n(CCCCC)c1)Cc3cccc(OC)c3</chem>	07.05
cb2_157	<chem>Oc1cc(CCCCC)cc2c1-c3cc(ccc3C(O2)(C)C)C</chem>	07.02
cb2_158	<chem>Fc1ccc(Cn2cc(c3ccccc32)C(=O)c4cccc5ccccc54)cc1</chem>	8.87
cb2_159	<chem>Fc1ccc(Cn2cc(c3ccccc32)C(=O)Nc4cccc5ccccc54)cc1</chem>	7.19
cb2_160	<chem>Fc1ccc(Cn2c3ccccc3c(n2)C(=O)N[C@@H](C(C)(C)C)C(=O)N)cc1</chem>	9.47
cb2_161	<chem>O=C(N[C@@H](Cc1ccccc1)C(=O)N)c2c3ccccc3n(n2)CC4CCCC4</chem>	8.36
cb2_162	<chem>O=C(c1c2ccccc2n(c1C)CCC)c3ccc(c4ccccc43)C</chem>	7.85
cb2_163	<chem>Clc1cccc(CC(=O)c2c3ccccc3n(CCCCC)c2)c1</chem>	6.97
cb2_164	<chem>Fc1cccc1CC(=O)c2c3ccccc3n(CCCCC)c2</chem>	7.41
cb2_165	<chem>FCCCCn1cc(c2ccccc21)C(=O)Nc3ccc4ccccc4c3</chem>	6.65
cb2_166	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(c4ccccc43)CCC</chem>	8.96
cb2_167	<chem>O=C(N)[C@@H](NC(=O)c1c2ccccc2n(n1)CC3CCCC3)C(C)C</chem>	9.35
cb2_168	<chem>O=C(c1cn(c2ccccc21)CCC)c3ccc(CCCC)c4ccccc34</chem>	7.28
cb2_169	<chem>Fc1ccc(Cn2cc(c3ccccc32)C(Oc4cccc5ccccc45)=O)cc1</chem>	8.61
cb2_170	<chem>FC(F)(F)CCCCc1cc(O)c2c(OC([C@@H]3CC=C(C[C@@H]23)C)(C)C)c1</chem>	7.52
cb2_171	<chem>FCCCCn1c2ccccc2c(n1)C(=O)N[C@@H](C(C)(C)C)C(=O)N</chem>	9.16
cb2_172	<chem>O=C(NC12CC3CC(C1)CC(C2)C3)c4c5ccccc5n(CCCCC)c4</chem>	8.91
cb2_173	<chem>O=C(c1cn(CCCCC#N)c2ccccc21)c3cccc4ccccc43</chem>	8.83
cb2_174	<chem>O=C(c1cn(CCCCC)c2ccccc21)c3ccc(OCC)c4ccccc43</chem>	7.98
cb2_175	<chem>FCCCCn1c(nc2ccccc21)C(=O)c3cccc4ccccc43</chem>	7.63