



UEPB

**UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I – CAMPINA GRANDE
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
DEPARTAMENTO DE FARMÁCIA
CURSO DE GRADUAÇÃO EM FARMÁCIA**

HILTHON ALVES RAMOS

**CONSTRUÇÃO DE MODELOS QSAR UTILIZANDO REGRESSÃO POR MÍNIMOS
QUADRADOS PARCIAIS E SELEÇÃO DE PREDITORES ORDENADOS A
PARTIR DE DERIVADOS ESPIRO-ACRIDÍNICOS COM ATIVIDADE ANTICÂNCER**

**CAMPINA GRANDE - PB
2022**

HILTHON ALVES RAMOS

CONSTRUÇÃO DE MODELOS QSAR UTILIZANDO REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS E SELEÇÃO DE PREDITORES ORDENADOS A PARTIR DE DERIVADOS ESPIRO-ACRIDÍNICOS COM ATIVIDADE ANTICÂNCER

Trabalho de Conclusão de Curso ou apresentada ao Departamento de Farmácia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Farmácia.

Área de concentração: Química Medicinal.

Orientador: Prof. Dr. José Germano Vêras Neto.

Coorientador: Prof. Dr. Ricardo Olimpio de Moura

**CAMPINA GRANDE
2022**

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

R175c Ramos, Hilthon Alves.

Construção de modelos QSAR utilizando regressão por mínimos quadrados parciais e seleção de preditores ordenados a partir de derivados espiro-acridínicos com atividade anticâncer [manuscrito] / Hilthon Alves Ramos. - 2022.

45 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Farmácia) - Universidade Estadual da Paraíba, Centro de Ciências Biológicas e da Saúde, 2022.

"Orientação : Prof. Dr. José Germano Vêras Neto , Departamento de Química - CCT."

"Coorientação: Prof. Dr. Ricardo Olimpio de Moura , Departamento de Farmácia - CCBS."

1. QSAR. 2. Derivados espiro-acridínicos. 3. Câncer de cólon. I. Título

21 ed. CDD.610

HILTHON ALVES RAMOS

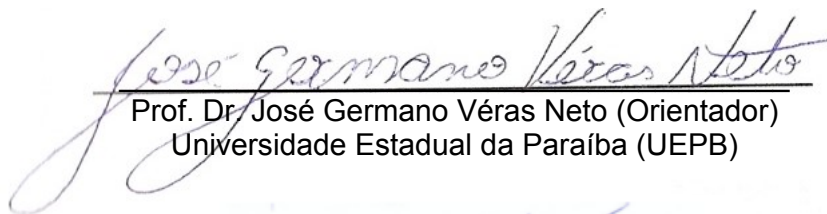
CONSTRUÇÃO DE MODELOS QSAR UTILIZANDO REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS E SELEÇÃO DE PREDITORES ORDENADOS A PARTIR DE DERIVADOS ESPIRO-ACRIDÍNICOS COM ATIVIDADE ANTICÂNCER

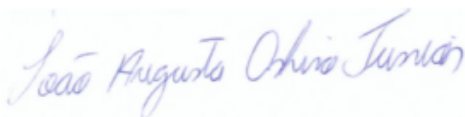
Trabalho de Conclusão de Curso ou apresentada ao Departamento de Farmácia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Farmácia.

Área de concentração: Química Medicinal.

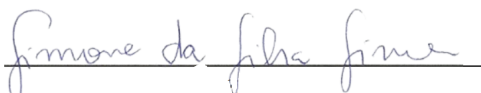
Aprovada em: _17/_06/_2022_.

BANCA EXAMINADORA


Prof. Dr. José Germano Vêras Neto (Orientador)
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. João Augusto Oshiro Júnior
Universidade Estadual da Paraíba (UEPB)



Profa. Dra. Simone da Silva Simões
Universidade Estadual da Paraíba (UEPB)

À minha mãe, pelo amor, incentivo e por toda ajuda para realização desse sonho, DEDICO.

AGRADECIMENTOS

À Deus, pela vida, pelo discernimento e por ter me dado a capacidade intelectual de concluir este trabalho.

À minha mãe, Laura Borges Alves e a minha tia, Inez Borges Alves por sempre se dedicarem e me apoiarem nas horas que precisei, eternamente agradecido.

Ao Prof. Dr. José Germano Vêras Neto (meu pai científico) e a Prof. Dra. Ana Claudia Dantas de Medeiros pela confiança depositada, por todos os ensinamentos, pelas oportunidades geradas e principalmente pela construção de uma amizade verdadeira. Foi a partir de vocês que decidi trilhar o caminho que sigo hoje, a pesquisa científica.

Ao Prof. Dr. Felipe Hugo Alencar Fernandes e ao Prof. Dr. João Augusto Oshiro Júnior por todos os ensinamentos compartilhados para me tornar um jovem pesquisador em busca de sabedoria.

À minha namorada e amiga, Cristhenes, por todo incentivo e apoio por ter estado ao meu lado durante todos os momentos durante essa fase final da graduação.

Ao Prof. Dr. Ricardo Olimpio de Moura pelo auxílio, parceria e acreditação em ceder seus dados para realização deste trabalho. Sem eles, essa pesquisa não seria possível.

Aos meus amigos da graduação, Beatriz Maria, Lucas, Carlos, Avyner e Moisés, pelos momentos de estudos, pelas brincadeiras e pela amizade sincera.

A todos os amigos do Laboratório de Química Analítica e Quimiometria (LQAQ), Vitor, Danúbio, Mariana Calixto, Rômulo, Maria Eduarda, Emilly, Gisele, Lêda, Mirelly, Ana Flávia e Paulo pelos momentos únicos covidados e pela amizade sincera.

A todos os amigos do Laboratório de Desenvolvimento e Ensaio de Medicamentos (LABDEM), Naara, Mariana Dantas, João Victor, Jessé, Kilma e Vimerson.

Aos professores do curso de Farmácia pelo incentivo e compreensão.

Aos funcionários da UEPB, em especial Ronald por todo apoio e auxílio na resolução de problemas durante a graduação.

A todos que de forma direta ou indireta fizeram parte dessa conquista.

[...] a verdadeira pobreza não é material e sim a solidão física e espiritual. [...]
Todos nós necessitamos dos demais para construir-nos. Ninguém se constrói sozinho. Necessitamos das demais pessoas para tecer as relações sociais, afetivas e emocionais que nos permitam existir. A mais simples das relações sociais é interpessoal. O “eu” que nos possibilita situar-nos na sociedade é também o “tu” com o qual os outros nos nomeiam.
(José Marín, 2020).

RESUMO

O câncer de cólon é o adenocarcinoma mais diagnosticado em homens, com uma alta taxa de mortalidade. Apesar de existirem diferentes terapias e alternativas, os pacientes geralmente preferem a quimioterapia. Entretanto, os medicamentos anticancerígenos apresentam graves efeitos colaterais, o que obriga a procurar por um medicamento mais seguro. As enzimas topoisomerase I e II estão entre os principais alvos terapêuticos do câncer, uma vez que são responsáveis pela diminuição da tensão gerada pelas supertorções do DNA, o que acarreta grande importância fisiológica. Os derivados de acridina suprimem a atividade dessas enzimas interrompendo o reparo e a replicação do DNA, o que induz a morte celular. Os estudos de QSAR (Relação Quantitativa entre Estrutura e Atividade) surgem como alternativa para identificar novos compostos com boa atividade biológica antes de sintetizá-los no laboratório e analisá-los *in vitro*. Neste sentido, o objetivo deste trabalho foi avaliar a relação entre moléculas de derivados espiro-acridínicos com potencial para atividade anticâncer a partir de estudos QSAR, utilizando algoritmos de regressão e seleção de variáveis. Foi utilizado um conjunto de 21 derivados espiro-acridínicos previamente sintetizados e avaliados preliminarmente frente a células de câncer de cólon divididos em três grupos: ACMD, AMTAC e a mistura dos dois. Cálculos semi-empíricos AM1 e de descritores moleculares 1D e 2D a partir do software PaDEL foram executados para construir o conjunto de preditores (variáveis) no estudo. O modelo quimiométrico apresentou baixo resíduo e ausência de amostras anômalas. Para os três grupos, os melhores resultados foram: ACMD (PCs = 1, RMSEC = 1.15, $r^2 = 0.9976$, $Q^2_{LOO} = 0.9976$), AMTAC (PCs = 2 e RMSEC = 3.91, $r^2 = 0.9519$ e $Q^2_{LOO} = 0.8729$) GERAL (PCs = 4 e RMSEC = 1.63, $r^2 = 0.9945$ e $Q^2_{LOO} = 0.9945$). O modelo de previsão foi proposto a partir de 21 novas moléculas para avaliação da atividade biológica. Duas moléculas para o modelo AMTAC (AMTACA-02 (70.54%) e AMTACB-04 (73.44%)) GERAL (AMTAC-01 (84.46%) e AMTACB-04 (76.21%)). Neste sentido, os modelos podem ser utilizados por pesquisadores que desejam sintetizar e avaliar novos compostos semelhantes às estruturas dos derivados espiro-acridínicos para o planejamento de novas moléculas com alta atividade contra o câncer de cólon.

Palavras-Chave: QSAR-2D; Derivados espiro-acridínicos; Câncer de cólon.

ABSTRACT

Colon cancer is the most commonly diagnosed adenocarcinoma in men, with a high mortality rate. Although different therapies and alternatives exist, patients generally prefer chemotherapy. However, anticancer drugs have serious side effects, which forces the search for a safer drug. The enzymes topoisomerase I and II are among the main therapeutic targets in cancer, since they are responsible for reducing the tension generated by DNA supertorsions, which is of great physiological importance. Acridine derivatives suppress the activity of these enzymes by interrupting DNA repair and replication, which induces cell death. QSAR (Quantitative Structure-Activity Relationship) studies emerge as an alternative to identify new compounds with good biological activity before synthesizing them in the laboratory and analyzing them in vitro. In this sense, the objective of this work was to evaluate the relationship between molecules of spiro-acridinic derivatives with potential anticancer activity from QSAR studies, using regression algorithms and variable selection. A set of 21 spiro-acridinium derivatives previously synthesized and preliminarily evaluated against colon cancer cells divided into three groups was used: ACMD, AMTAC and a mixture of the two. Semi-empirical AM1 and 1D and 2D molecular descriptor calculations from PaDEL software were performed to build the set of predictors (variables) in the study. The chemometric model showed low residuals and no anomalous samples. For the three groups, the best results were: ACMD (PCs = 1, RMSEC = 1.15, $r^2 = 0.9976$, $Q^2_{LOO} = 0.9976$), AMTAC (PCs = 2 and RMSEC = 3.91, $r^2 = 0.9519$ and $Q^2_{LOO} = 0.8729$) GENERAL (PCs = 4 and RMSEC = 1.63, $r^2 = 0.9945$ and $Q^2_{LOO} = 0.9945$). The prediction model was proposed from 21 new molecules for biological activity evaluation. Two molecules for the AMTAC model (AMTACA-02 (70.54%) and AMTACB-04 (73.44%)) GENERAL (AMTAC-01 (84.46%) and AMTACB-04 (76.21%)). In this sense, the models can be used by researchers who wish to synthesize and evaluate new compounds similar to the structures of spiro-acridinic derivatives for the planning of new molecules with high activity against colon cancer.

Keywords: 2D-QSAR; Spiro-acridine derivatives; Colon cancer.

LISTA DE ILUSTRAÇÕES

Figura 1 – Descobertas pioneiras que levaram à evolução do métodos de QSAR.....	16
Figura 2 – Esquema geral de etapas de seleção de variáveis usando o algoritmo OPS.....	24
Figura 3 – Estruturas químicas dos derivados espiro-acridínicos do grupo ACMD com atividade anticâncer contra linhagens de células humanas cancerosas HCT-116.....	25
Figura 4 – Estruturas químicas dos derivados espiro-acridínicos do grupo AMTAC com atividade anticâncer contra linhagens de células humanas cancerosas HCT-116.....	26
Figura 5 – Fluxograma da pesquisa.....	28
Figura 6 – Gráficos de Escores da PCA do grupo ACMD.....	31
Figura 7 – Gráficos de Influência da PCA do grupo ACMD.....	31
Figura 8 – Gráficos de Escores da PCA do grupo AMTAC.....	31
Figura 9 – Gráficos de Influência da PCA do grupo AMTAC.....	32
Figura 10 – Gráficos de Escores da PCA do grupo GERAL.....	32
Figura 11 – Gráficos de Influência da PCA do grupo GERAL.....	32

LISTA DE TABELAS

Tabela 1 – Atividade biológica das moléculas dos grupos AMCD e AMTAC com atividade anticâncer com dosagem de 50 µM contra linhagens de células humanas (HCT-116) para câncer de cólon.....	29
Tabela 2 – Resultados da seleção de variáveis a priori para os modelos AMCD, AMTAC e GERAL.....	30
Tabela 3 – Seleção de variáveis OPS-PLS para os modelos ACMD, AMTAC e GERAL.....	33
Tabela 4 – Figuras de mérito para o conjunto de calibração de regressão PLS para os modelos ACMD, AMTAC e GERAL.....	34
Tabela 5 – Valores medidos vs preditos e desvios padrões para os modelos ACMD, AMTAC e GERAL.....	35
Tabela 6 – Figuras de mérito para o conjunto de validação da seleção de variáveis OPS-PLS para os modelos ACMD, AMTAC e GERAL.....	36
Tabela 7 – Valores previstos das moléculas do modelo ACMD.....	38
Tabela 8 – Valores previsto das moléculas do modelo AMTAC.....	38
Tabela 9 – Valores previstos das moléculas do modelo GERAL.....	39

LISTA DE ABREVIATURAS E SIGLAS

MLR	Regressão linear múltipla (do inglês, <i>Multiple Linear Regression</i>)
NIPALS	Mínimos Quadrados Parciais Não Lineares (do inglês, <i>Non Linear Iterative Partial Least Squares</i>)
OLS	Mínimos quadrados ordinários (do inglês, <i>Ordinary Least Squares</i>)
OPS	Seleção ordenada de preditores (do inglês, <i>Ordered Predictor Selection</i>)
PCA	Análise de Componentes Principais (do inglês, <i>Principal Component Analysis</i>)
PCR	Regressão por Componentes Principais (do inglês, <i>Principal Component Regression</i>)
PLSR	Regressão por Mínimos Quadrados Parciais (do inglês, <i>Partial Least Squares Regression</i>)
QSAR	Relação Estrutura-Atividade Quantitativa (do inglês, <i>Quantitative structure–activity relationship</i>)
Q^2_{Loo}	Coeficiente de Correlação em Validação Cruzada Completa (do inglês, <i>Leave-one-out Cross-validated Correlation Coefficient</i>)
RMSEC	Raiz do Erro Médio Quadrático de Calibração (do inglês, <i>Root Mean Square Error of Calibration</i>)
RMSECV	Raiz do Erro Médio Quadrático de Validação Cruzada (do inglês, <i>Root Mean Square Error of Cross Validation</i>)
RMSEP	Raiz do Erro Médio Quadrático de Predição (do inglês, <i>Root Mean Squared Error of Prediction</i>)
SPXY	Particionamento do Conjunto de Amostras com Base na Distância conjunta x - y (do inglês, <i>Sample Set Partioning Based on Joint x-y Distance</i>)
VL	Variáveis Latentes

SUMÁRIO

1 INTRODUÇÃO.....	12
2 OBJETIVOS.....	14
2.1 Objetivo geral.....	14
2.2 Objetivos específicos.....	14
3 FUNDAMENTAÇÃO TEÓRICA	15
3.1 QSAR.....	15
3.1.2 QSAR-2D	17
3.1.3 Descritores moleculares.....	18
3.2 Quimiometria.....	19
3.2.1 Calibração multivariada.....	20
3.2.2 <i>Regressão por mínimos quadrados parciais (PLS)</i>	21
3.2.3 Seleção de Variáveis	22
3.2.4 <i>Seleção Ordenada de Preditores – OPS</i>	23
3.3 Derivados espiro-acridínicos e QSAR-2D.....	24
4 METODOLOGIA	25
4.1 Otimização geométrica e cálculo dos descritores moleculares	27
4.2 Avaliação biológica.....	27
4.3 Desenvolvimento de métodos quimiométricos	27
5 RESULTADOS E DISCUSSÕES	29
5.1 Modelo biológico	29
5.2 Seleção de variáveis a priori	29
5.2.1 Análise exploratória.....	30
5.3 Análise dos modelos OPS-PLS	33
5.4 Validação dos modelos	35
5.5 Predição de novas moléculas como possíveis candidatos a fármacos	36
6 CONSIDERAÇÕES FINAIS.....	40
7 PERSPECTIVAS	41
REFERÊNCIAS	42

1 INTRODUÇÃO

No mundo, a segunda maior causa de mortes é devida ao câncer, atrás apenas das doenças cardiovasculares. Cerca de 19,3 milhões de novos casos e quase 10 milhões de mortes pela mesma doença ocorreram em 2020 em nível mundial (WHO, 2020; SUNG *et al.*, 2021). No Brasil, os números são de 309.750 novos casos e 121.686 mortes por ano (INCA, 2020). Em termos da incidência primária de tumores no Brasil, entre homens e mulheres, o câncer de cólon é segundo de maior prevalência.

Visando a redução do crescimento de células tumorais ou sua completa destruição, destaca-se o tratamento quimioterápico, que utiliza compostos químicos (agentes quimioterápicos) no combate as enfermidades. No entanto, sua eficácia é limitada pelo processo de resistência e toxicidade não seletiva. Dentre as opções de quimioterápicos, temos os derivados de acridina que têm sido investigados e utilizados para o tratamento do câncer por mais de um século e vem sendo testado como intercaladores de DNA, pois a citotoxicidade da maioria das drogas à base de acridina é baseada na sua capacidade de intercalar o DNA e suprimir a atividade da topoisomerase I e II, que estão entre os principais alvos biológicos na terapia do câncer. Sendo assim, é importante o desenvolvimento de novos análogos por meio do estudo da atividade biológica e dos intercaladores de DNA, o que irá proporcionar a descoberta de moléculas mais seguras, como também seus respectivos mecanismos de ação (SHARMA *et al.*, 2020).

Na tentativa de obter novos intercaladores de DNA mais seletivos, Pinheiro Segundo (2020) e Gouveia (2017) inovaram desenvolvendo uma série de derivados espiro-acridínicos para combater o câncer de cólon. Todavia, estudos envolvendo modelos que correlacionam quantitativamente a estrutura e atividade biológica das moléculas ainda não foram realizados, o que iria permitir desenvolver compostos com propriedades químicas, físicas e biológicas desejadas.

As abordagens tradicionais no planejamento de novas moléculas envolvem a sínteses, avaliação estrutural e de rota, o que torna o processo extremamente caro e demorado. Para contornar estes inconvenientes, propõe-se o uso de modelos que relacionam quantitativamente a estrutura química e atividade biológica (QSAR). Este método permite produzir e testar hipóteses para facilitar a compreensão das

interações entre as moléculas e acelerar o processo de descoberta de candidatos a fármacos de maneira econômica, permitindo, assim, sintetizar apenas moléculas que apresentem uma alta atividade biológica (MURATOV *et al.*, 2020).

Nesta perspectiva, o presente trabalho surge como alternativa para obter modelos QSAR para a série de derivados espiro-acridínicos. Para seu desenvolvimento, foi utilizado a regressão de mínimos quadrados parciais (PLS) e a seleção ordenada de preditores (OPS).

2 OBJETIVOS

2.1 Objetivo geral

- Construir modelos quantitativos de relação estrutura e atividade (QSAR) para derivados espiro-acridínicos com potencial atividade anticâncer de cólon e gerar modelos capazes de prever a atividade de compostos ainda não testados.

2.2 Objetivos específicos

- Selecionar derivados espiro-acridínicos em base de dados na literatura;
- Calcular a geometria molecular e otimizar cada composto do estudo através de métodos semi-empíricos;
- Identificar propriedades físico-químicas que são essenciais para atividade dos compostos contra o câncer de cólon;
- Validar os modelos usando os critérios das figuras de méritos apropriadas.

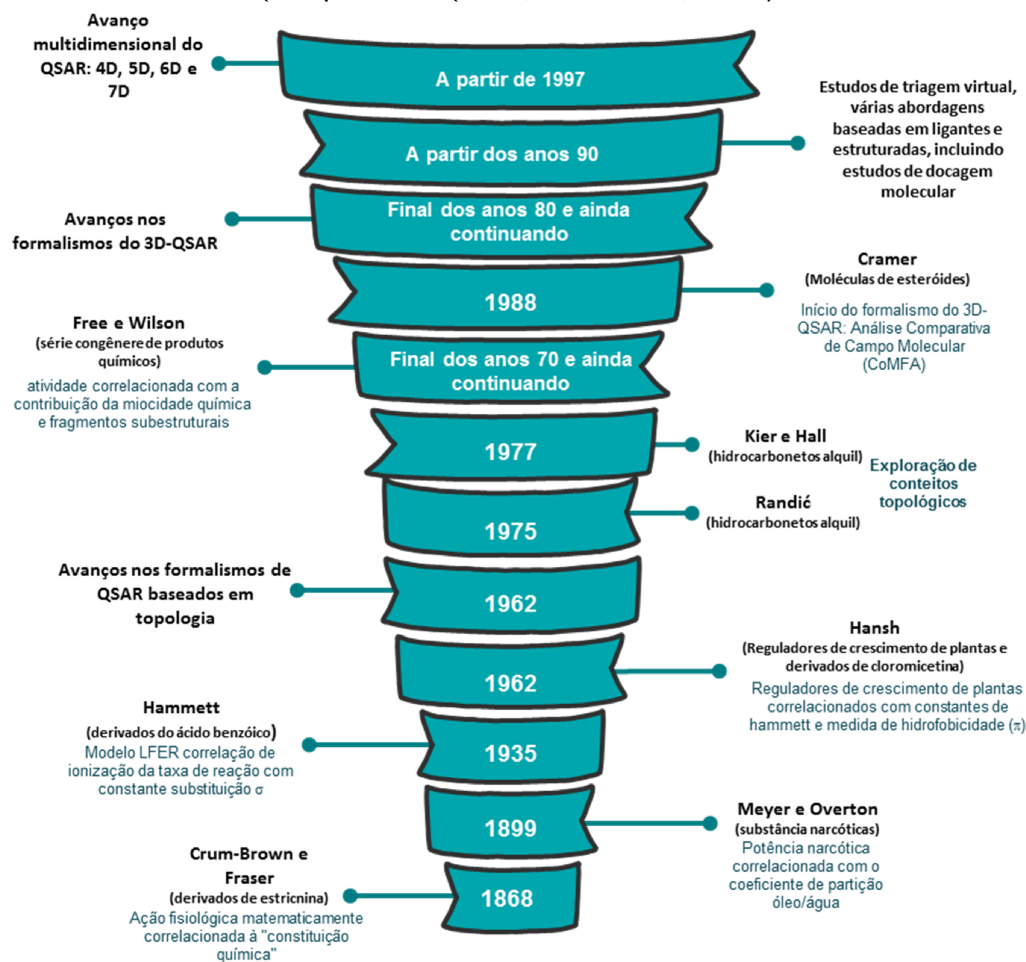
3 FUNDAMENTAÇÃO TEÓRICA

3.1 QSAR

A modelagem da relação quantitativa entre estrutura química e atividade biológica (QSAR, do inglês “Quantitative structure-activity relationship”) é uma área da química medicinal. Seu surgimento inicia com Mendeleev, que pode ser citado como um dos primeiros cientistas no campo da química que utilizou o conceito de correlação química em 1870, formulando a regra dos oito. Acredita-se que a ideologia do QSAR surgiu no campo da toxicologia e mais tarde foi apoiada por experimentos em química e física. Entretanto, Crum-Brown e Fraser são considerados como os pioneiros no reino da modelagem QSAR que representavam a ação fisiológica em termos de "constituição química" em 1868, embora esta frase não seja um conceito bem explicado naquela época. Assim, podemos ver que as observações iniciais da análise do QSAR derivaram de estudos toxicológicos e os parâmetros correlatos eram principalmente atributos físico-químicos. O estudo do desenvolvimento de descritores quantitativos para modelos de correlação matemática foi colocado em evidência por Hammett, que introduziu a medida constante de substituição eletrônica decisiva Hammett sigma (σ) ao relacionar a taxa de reação relativa de derivados de ácido benzóico meta e para-substituído (LIN; LI; LIN, 2020; MURATOV *et al.*, 2020).

Após o relato do surgimento do QSAR, a contribuição pioneira para área foi o desenvolvimento da equação de Hansch no início dos anos 60. Corwin Hansch recebeu o título de "Pai do QSAR moderno", que realizou estudos sobre reguladores de crescimento de plantas usando a medida de hidrofobicidade relativa de substituto (π). A forma linear da equação foi concebida por Hansch e Fujita através da incorporação do termo constante Hammett. No final da década de 1960, os estudos sobre a teoria dos índices gráficos topológicos envolvendo conceitos de matemática e química levaram ao desenvolvimento de descritores quantitativos sobre uma base puramente teórica. Este estudo abriu uma nova possibilidade no campo da química teórica, especialmente com referência ao formalismo QSAR que, posteriormente levou ao desenvolvimento do método da diferença topológica mínima (MTD) de Simon, parâmetros de índice de conectividade de Randić, Kier e Hall, entre outros (JIMÉNEZ-LUNA *et al.*, 2021). A Figura 1 resume as conquistas pioneiras que levaram à evolução histórica do formalismo do QSAR (MURATOV *et al.*, 2020).

Figura 1 - Descobertas pioneiras que levaram à evolução do métodos de QSAR (Adaptado de (ROY; KAR; DAS, 2015)).



Fonte: Elaborada pelo autor, 2022.

Considerando as bases matemáticas envolvidas na quantificação das informações químicas, os descritores podem apresentar a dimensão da análise QSAR correspondente. Como a extração de informações químicas envolve várias suposições hipotéticas, o estudo QSAR pode ser visto de uma perspectiva dimensional. No quadro 1, é mostrado os métodos de QSAR utilizando informações químicas de dimensões variáveis.

Quadro 1 - Perspectiva dimensional dos métodos de QSAR (Adaptado de (ROY; KAR; DAS, 2015))

Perspectiva dimensional do QSAR	
QSAR-0D	Fórmula química derivada
QSAR-1D	Fragmento subestrutural derivado
QSAR-2D	Teoria derivada do índice gráfico topológico
QSAR-3D	Geometria espacial derivada
QSAR-4D	Conformação, orientação e representação do estado de protonação derivada
QSAR-5D	Parâmetros de ajuste induzido derivados (modelo virtual ou pseudo receptor baseado em ligante)
QSAR-6D	QSAR-5D + outras condições de solvatação derivadas
QSAR-7D	Dados reais do modelo de receptor baseado em alvos derivados

Fonte: Elaborada pelo autor, 2022.

Dentre os vários métodos de QSAR, a proposta deste trabalho será usar estudos de relações quantitativas entre estrutura e atividade bidimensionais (QSAR-2D).

3.1.2 QSAR-2D

O estudo do QSAR-2D ou QSAR clássico é o mais utilizado dentre os métodos de relação quantitativa estrutura e atividade. Baseia-se na construção de diagramas que relacionam as propriedades moleculares analisadas em uma distribuição cartesiana. Faz uso de parâmetros físico-químicos (lipofílico, estérico e eletrônico) e estruturais linearmente relacionados para construir o modelo, seguindo a seguinte equação para modelos do tipo multidimensionais (DUARTE *et al.*, 2020).

$$Y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots + a_n x_n \quad (5)$$

Onde:

- Y = variáveis dependentes (atividade, propriedade, dentre outros)
- X_1, X_2, X_n = variáveis independentes (características estruturais ou físico-químicas)
- a_1, a_2, a_n = contribuições individuais dos descritores individuais

Para o desenvolvimento do QSAR-2D, vários pré-requisitos são necessários para o desenvolvimento dos modelos, como: (1) os compostos a serem estudados devem ser congêneres estreitamente relacionados, aumentando assim a probabilidade de ter o mesmo mecanismo de ação; (2) os dados da atividade biológica a serem usado na modelagem deve ser preciso e medido sob condições uniformes e (3) o parâmetro de atividade deve ser intrinsecamente aditivo. Para a confiabilidade estatística de tais modelos, é desejável ter uma alta relação entre o número de observações e o número de termos desconhecidos nas equações lineares (JIMÉNEZ-LUNA *et al.*, 2021).

3.1.3 Descritores moleculares

Os descritores moleculares são parâmetros quantitativos calculados experimentalmente ou computacionalmente, fornecendo informações específicas de uma determinada molécula estudada. De acordo com o método do cálculo e sua determinação, os descritores podem variar desde índices constitucionais, impressões digitais, propriedades físico-químicas, estruturais, topológicas, eletrônicos, espaciais entre outros. Os descritores também podem ser classificados de acordo com sua dimensão, como mostrado no quadro abaixo (MAHALAKSHMI; JAHNAVI, 2020).

Quadro 2 - Diferentes descritores empregados no estudo de QSAR com base em sua dimensão (Adaptado de (ROY; KAR; DAS, 2015))

Dimensão dos descritores	Parâmetros
Descritores-0D	Índices constitucionais, propriedade molecular, átomos e contagem de ligações
Descritores-1D	Contagem de fragmentos e impressões digitais
Descritores-2D	Parâmetros topológicos, estruturais, físico-químicos e descritores termodinâmicos

Descritores-3D

Parâmetros eletrônicos, espaciais, análise da forma molecular, análise do campo molecular e análise de superfície de receptor

Fonte: Elaborada pelo autor, 2022.

No QSAR-2D, destacam-se os descritores físico-químicos e topológicos como os mais utilizados. Os descritores topológicos são calculados com base na representação gráfica de moléculas e, portanto, não exigem a estimativa de quaisquer propriedades físico-químicas nem precisam dos cálculos rigorosos envolvidos na derivação dos descritores químicos. A representação da estrutura da molécula depende de sua topologia gráfica, o que indica a posição dos átomos individuais e o conexões entre eles. Os parâmetros físico-químicos são projetados com base em propriedades físicas e químicas das moléculas. Estes parâmetros estão interligados com as propriedades farmacocinéticas dos fármacos, o que poderá alterar seu comportamento bioativo. Dentre os parâmetros mais comuns que afetam a bioatividade dos compostos estão a hidrofobicidade, caráter eletrônico e estérico. (WANG *et al.*, 2021).

Nesta perspectiva, a resposta obtida do modelo QSAR está diretamente correlacionada com sua dimensionalidade, ou seja, quanto maior a sua dimensionalidade, maior será o leque de parâmetros que são levadas em consideração. Como etapa prévia, os estudos de QSAR-2D são utilizados pela química medicinal e computacional, uma vez que seu desenvolvimento do modelo é rápido e fácil em relação aos outros tipos de modelos com maiores dimensões. (FERREIRA, 2015).

Portanto, existem milhares de descritores disponíveis que podem ser calculados de forma virtual. Todavia, grande parte dos descritores calculados apresentam informação contraditória ou mesmo redundantes, o que dificulta encontrar uma explicação química para o modelo criado. Nesta perspectiva, a quimiometria é uma alternativa viável que faz uso de métodos de seleção de variáveis para expressar as variáveis ditas importantes para construção do modelo, o que consequentemente aumenta sua capacidade preditiva (DE OLIVEIRA, 2021).

3.2 Quimiometria

O grande avanço na área instrumental ocorrido nos últimos 50 anos atingiu todas as áreas da ciência e especialmente a química. A criação de espectrofotômetros

e cromatógrafos alavancaram a rotina dos laboratórios analíticos. Entretanto, tais instrumentos produzem uma grande quantidade de dados numéricos e com difícil interpretação, necessitando-se de métodos que pudessem transformar esses dados numéricos em informação química. Sendo assim, a quimiometria surge como ferramenta matemática e estatística capaz de obter a máxima informação útil de um conjunto de dados. As principais áreas da quimiometria são: processamento de sinais analíticos, planejamento de otimização de experimentos, calibração multivariada, métodos de reconhecimento de padrões e entre outros. (BYSTRZANOWSKA; TOBISZEWSKI, 2020).

O presente estudo irá utilizar métodos de calibração multivariada a partir dos algoritmos de regressão, o que irá promover a quantificação da atividade biológica dos compostos bioativos.

3.2.1 Calibração multivariada

Um modelo de calibração é formado a partir de uma variável medida, como uma resposta instrumental, em uma variável química informativa, ou seja, o analito alvo. A partir disso, é possível obter uma relação matemática entre as respostas do analito e do instrumento pela medição de amostras contendo o analito em diferentes concentrações conhecidas (MISHRA *et al.*, 2021).

A calibração univariada é aplicada quando um padrão puro está disponível e uma série de padrões de calibração podem ser preparadas, como o exemplo clássico da construção de uma curva analítica. As amostras de calibração são medidas em um único comprimento de onda para dados espectroscópicos, e os mínimos quadrados ordinários (OLS, do inglês “Ordinary Least Squares”) são calculados para estimar o coeficiente de regressão do modelo de calibração. Uma desvantagem dos modelos univariados é que um único comprimento de onda é completamente seletivo para o analito, entretanto, pode ser que apenas um único ponto não seja suficiente para descrever de forma quantitativa o modelo. Já na calibração multivariada, os modelos são construídos utilizando uma faixa de vários comprimentos de onda relacionadas a um ou mais propriedades desconhecidas das amostras, o que atribui mais informação e robustez ao mesmo (OLIVERI; MALEGORI; CASALE, 2020).

Os modelos de calibração multivariadas são construídos em forma de matriz a partir dos *scores* e *loadings*, em que a mesma possui as respostas (variáveis independentes) denominada matriz **X** e a matriz **Y** contendo o parâmetro de referência

(variável dependente). Nos estudos de QSAR, as variáveis independentes são os descritores moleculares e a variável dependente é a atividade biológica. Outra etapa importante é capacidade preditiva do modelo, que pode ser avaliada por meio de validação interna ou externa. Para validação externa, utilizam-se amostras independentes externas ao modelo de origem conhecidas, assim podendo-se verificar a capacidade preditiva do modelo. Já a validação interna é utilizada geralmente quando o número de amostras é limitado. Um exemplo de modelo interno muito utilizado é a validação cruzada (do inglês, *cross-validation*), sendo muito usado em modelos PLS. A validação externa fundamenta-se em separar o conjunto de calibração em vários segmentos, removendo uma amostra por vez e recriando o modelo de calibração e testando com a amostra removida para validar o modelo. Dessa forma, a abordagem mais comum é a validação cruzada leave-one-out (LOO), onde ocorre seguindo a mesma metodologia citada anteriormente, cuja uma única amostra é retirada por vez e repetida até que todas as mesmas tenham sido deixadas de fora (DE OLIVEIRA, 2021; OLIVERI; MALEGORI; CASALE, 2020).

Nesta perspectiva, o constante desenvolvimento de novos algoritmos para construir modelos de calibração multivariada favorecem a resolução de problemas de vários tipos, como os dados espectrais, cromatográficos, métodos QSAR, onde o analito alvo (atividade biológica) é uma função de descritores moleculares e entre outros. Dentre os desafios de um modelo multivariado, estão a multicolinearidade e a deficiência de classificação dos dados. Sendo assim, uma ampla gama de algoritmos foram criados como a regressão por mínimos quadrados parciais (PLS, do inglês “Partial Least Square Regression”), regressão por componentes principais (PCR, do inglês “Principal Componentes Regression”) e regressão linear múltipla (MLR, do inglês “Multiple Linear Regression”) (CHIAPPINI *et al.*, 2020).

Para o estudo de QSAR proposto, o algoritmo PLS será utilizado, visto que possui vários pontos positivos, pois identifica fatores que melhor modelam a atividade biológica dos compostos, trabalha com eficiência com um conjunto de dados com descritores altamente correlacionados e que apresentarem alto resíduo, e lida de forma mais robusta utilizando mais descritores do que moléculas.

3.2.2 Regressão por mínimos quadrados parciais (PLS)

No ano de 1975 em estudos desenvolvidos por Wold *et al.* (1975) surge o algoritmo PLSR com uma modelagem de conjunto para dados complexos. A

modelagem PLS utiliza a informação contida numa matriz \mathbf{X} de dados e a matriz resposta \mathbf{Y} (atividade biológica), modelando uma máxima correlação entre esses dois conjunto de variáveis, o que resulta em novas variáveis chamadas de variáveis latentes, componentes ou fatores (CHIAPPINI *et al.*, 2020)

As matrizes \mathbf{X} e \mathbf{Y} são decompostas em produto com duas matrizes menores, chamadas de pesos e escores, respectivamente, como descrito nas Equações 1 e 2.

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (1)$$

$$\mathbf{X} = \mathbf{UQ}^T + \mathbf{F} \quad (2)$$

Para as Equações (1) e (2), \mathbf{T} e \mathbf{U} são as matrizes de *scores*, \mathbf{P} e \mathbf{Q} são as matrizes de loadings, e \mathbf{E} e \mathbf{F} são as matrizes de erro de \mathbf{X} e \mathbf{Y} respectivamente.

Dentre as várias formas de obter os parâmetros de um modelo PLS, o mais conhecido é o Algoritmo dos Mínimos Quadrados Parciais Iterativos Não-Lineares (NIPALS, do inglês “Nonlinear Iterative Partial Least Squares”), proposto por World. O modelo relaciona de uma forma linear os *scores* de uma matriz \mathbf{X} com os *scores* de uma matriz \mathbf{Y} , seguindo as equações 3 e 4, respectivamente (COOK; FORZANI, 2020).

$$\mathbf{U} = \mathbf{BT} + \mathbf{G} \quad (3)$$

$$\mathbf{X} = \mathbf{BTQ}^T + \mathbf{H} \quad (4)$$

Os termos nas Equações (3) e (4) são o coeficiente ajustado, \mathbf{B} , calculado geralmente usando o algoritmo NIPALS, \mathbf{G} é a matriz de resíduos dos *scores*, e \mathbf{H} a matriz de resíduos de concentração (COOK; FORZANI, 2020).

3.2.3 Seleção de Variáveis

Em um modelo multivariado, é comum ter uma grande quantidade de variáveis e poucas amostras, como nos estudos de QSAR. Desse modo, muitas dessas variáveis não contribuem para a construção do modelo, ocasionando principalmente problemas de multicolinearidade. Sendo assim, o objetivo dessas técnicas é selecionar um número reduzido de variáveis significantes afim de obter uma ótima resposta preditiva, a partir da remoção de variáveis não significativa e redundantes, o

que interfere nos modelos de regressão e reconhecimento de padrões (MEHMOOD; SÆBØ; LILAND, 2020). Um método que foi proposto recentemente na literatura é o OPS, cuja finalidade foi a resolução de problemas em estudos de QSAR, conseguindo produzir excelentes resultados nesses estudos (HALDER; DIAS SOEIRO CORDEIRO, 2020).

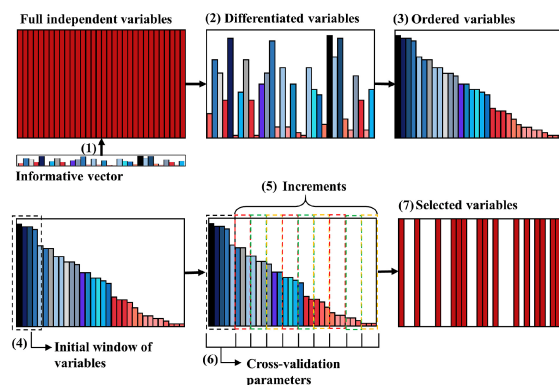
3.2.4 Seleção Ordenada de Preditores – OPS

A seleção de preditores ordenados usa vetores informativos que devem estar bem relacionados com a variável dependente (atividade biológica). Dessa forma, assume-se que os elementos de um vetor informativo cujo valores são altos contêm mais informação relevante da propriedade de interesse (variáveis dependente) (FERREIRA, 2015).

Na figura 2, é mostrato de forma intuitiva todo esquema realizado durante a aplicação do OPS, seguindo os seguintes passos (DE PAULO *et al.*, 2020):

- (1) São obtidos vetores que carregam a informação;
- (2) Ocorre a diferenciação dos valores absolutos dos vetores informativos na matriz de dados;
- (3) As variáveis são ordenadas de forma decrescente;
- (4) Um subconjunto de variáveis (janelas) é definido para construção do modelo;
- (5) Variáveis são adicionadas como forma de incremento para que uma percentagem de variáveis seja levada em conta;
- (6) Calculam-se os parâmetros de validação cruzada para cada modelo;
- (7) Os parâmetros de qualidade são calculados para validar o modelo para definição do melhor modelo.

Figura 2 - Esquema geral de etapas de seleção de variáveis usando o algoritmo OPS (ROQUE *et al.*, 2019)



Fonte: (ROQUE *et al.*, 2019).

3.3 Derivados espiro-acridínicos e QSAR-2D

As acridinas são compostos aromáticos heterocíclicos que possuem atividade anticancerígena por se ligarem fortemente ao DNA, inibindo as enzimas topoisomerase I e II, causando dano ao DNA, interrompendo o reparo e a replicação do DNA, o que leva a morte celular. Seguindo essa perspectiva, novos derivados espiro-acridínicos foram obtidos a partir da acridina por reações de condensação seguidas de ciclização espontânea. Esses compostos apresentaram capacidade de ligação ao DNA, inibição da enzima topoisomerase e atividade anticancerígena *in silico* e *in vitro* (DUARTE *et al.*, 2021; OYEDELE *et al.*, 2020).

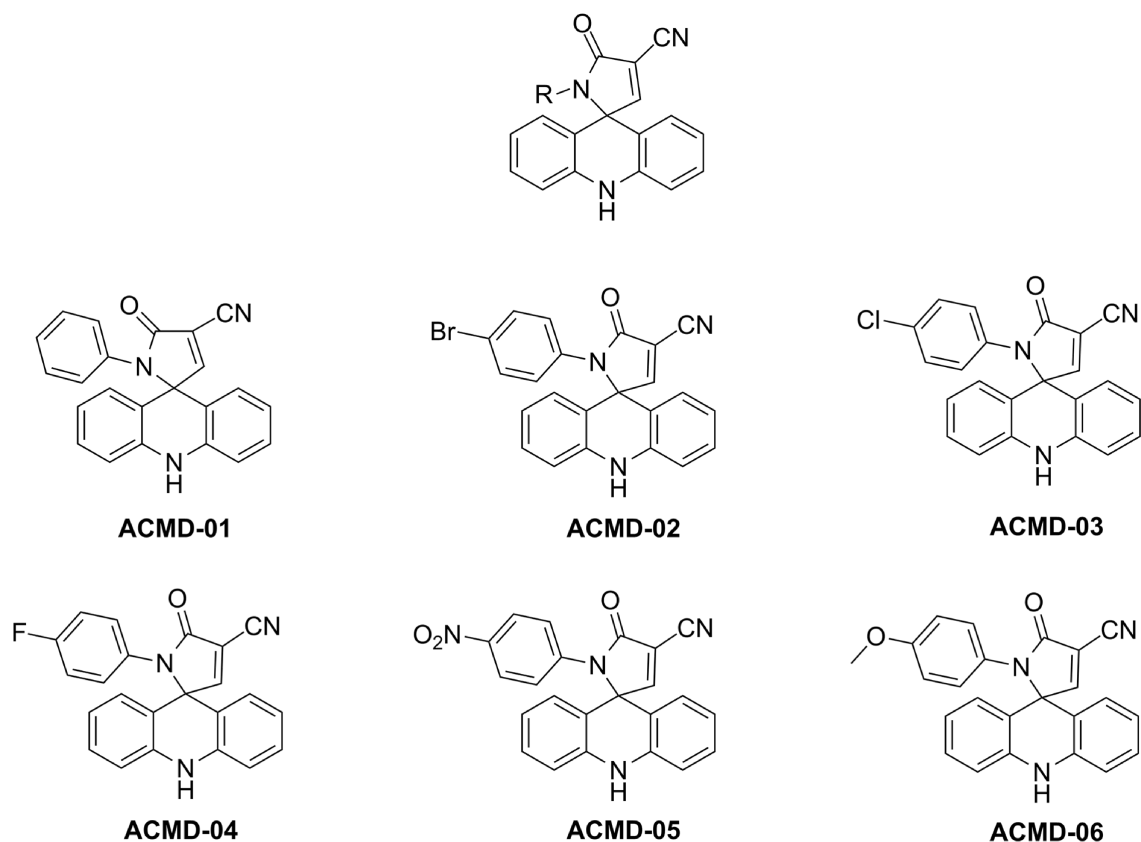
Os novos derivados espiro-acridínicos foram descritos na literatura recentemente (ALMEIDA *et al.*, 2016). A quantidade desses compostos que foram sintetizados e avaliados biologicamente ainda é um fator limitante para o desenvolvimento dos compostos. Como o processo de rota, síntese e avaliação biológica demandam muito tempo e custo, os métodos de QSAR surgem como alternativa para desenvolver novos compostos com a atividade biológica requerida pelos pesquisadores. Em relação a isso, estudos envolvendo QSAR, derivados espiro-acridínicos e câncer de cólon ainda são escassos na literatura.

Em uma pesquisa realizada no Portal de Periódicos da CAPES, não foi encontrado nenhum artigo utilizando QSAR-2D para prever a atividade anticâncer de cólon a partir de derivados espiro-acridínicos até o momento.

4 METODOLOGIA

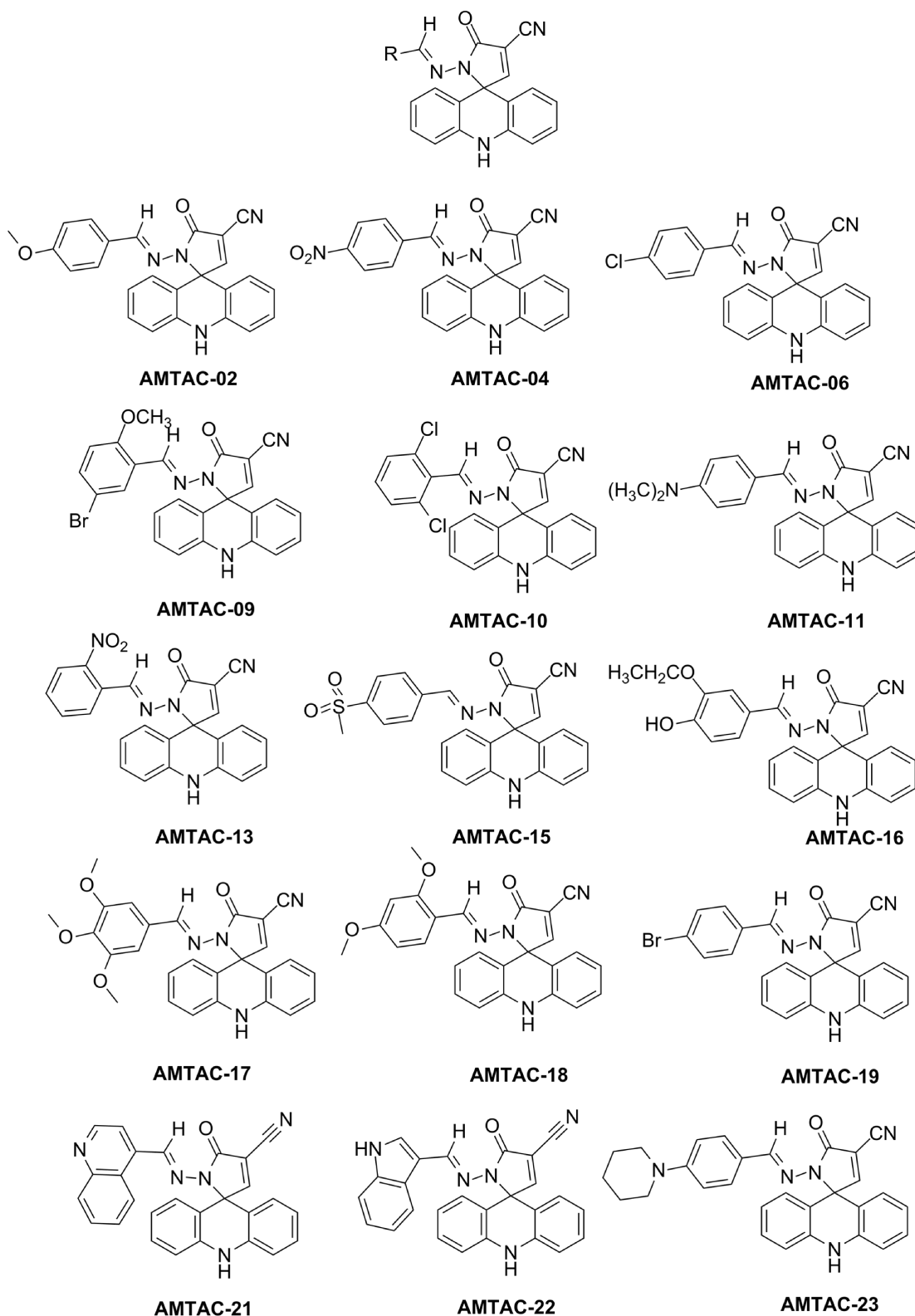
Foram utilizadas 21 moléculas para construção dos modelos, subdivididas em dois grupos: ACMD (6 moléculas) e AMTAC (15 moléculas). Tais moléculas foram sintetizadas e caracterizadas (PINHEIRO SEGUNDO, 2020; GOUVEIA, 2017). As moléculas escolhidas podem ser vistas nas figuras 3 e 4.

Figura 3 - Estruturas químicas dos derivados espiro-acridínicos do grupo ACMD com atividade anticâncer contra linhagens de células humanas cancerosas HCT-116



Fonte: Elaborada pelo autor, 2022.

Figura 4 - Estruturas químicas dos derivados espiro-acridínicos do grupo AMTAC com atividade anticâncer contra linhagens de células humanas cancerosas HCT-116



Fonte: Elaborada pelo autor, 2022.

4.1 Otimização geométrica e cálculo dos descritores moleculares

As estruturas das moléculas foram desenhadas utilizando os programas ChemDraw Ultra 12.0 e o ChemDraw 3D. Em seguida, as moléculas foram otimizadas utilizando o método semi-empírico Austin Model 1 (AM1) a partir do software MOPAC (Molecular Orbital PACKage).

Dentre a vasta gama de programas que permitem o cálculo de descritores moleculares, o PaDEL foi escolhido por ser um programa gratuito e de fácil acesso, obtendo resultados em um curto período de tempo. Atualmente, o software calcula 1.875 descritores (1.444 descritores 1D, 2D e 431 descritores 3D) e 12 tipos de impressões digitais. Foram calculados os descritores 1D e 2D para construção dos modelos.

4.2 Avaliação biológica

O ensaio para avaliar a atividade biológica não fez parte deste estudo. Os dados utilizados foram coletados conforme os trabalhos já publicados. A metodologia detalhada pode ser encontrada em Pinheiro Segundo (2020).

4.3 Desenvolvimento de métodos quimiométricos

Foram construídos três modelos: ACMD (6 amostras), AMTAC (15 amostras) e GERAL (6 amostras ACMD + 15 amostras AMTAC) para obtenção das matrizes completas com os descritores moleculares. O pré-processamento usado foi o *autoescalamento*.

A construção dos modelos iniciaram por meio de uma seleção de variáveis a priori e depois pelo corte de correlação de variáveis com significância abaixo de 30% em relação a atividade biológica. Em seguida, foi realizada uma análise exploratória para os três grupos utilizando PCA para avaliar o comportamento das moléculas em relação aos descritores moleculares. As matrizes foram divididas em conjunto de calibração e validação por meio do método de seleção de amostras, o SPXY, para os conjuntos AMTAC e GERAL. Em relação ao grupo ACMD, a seleção das amostras não foi aplicada devido ao grupo ter um número de amostras muito pequeno.

O método de seleção de variáveis OPS-PLS foi utilizado para construção dos modelos de regressão e a validação interna leave-one-out para validação dos modelos.

O OPS-PLS foi calculado utilizando o programa *QSAR modeling* (MARTINS; FERREIRA, 2013). Para o OPS-PLS, a seleção foi realizada até que fosse fornecido um subconjunto com o menor valor de RMSECV durante a validação cruzada para cada modelo. As condições utilizadas para o cálculo do OPS foram:

- ACMD (LV OPS = 6, LV_model = 1, Janelas = 1 e Incremento = 1)
- AMTAC (LV OPS = 12, LV_model = 2, Janelas = 1 e Incremento = 1)
- GERAL (LV OPS = 18, LV_model = 3, Janelas = 1 e Incremento = 1)

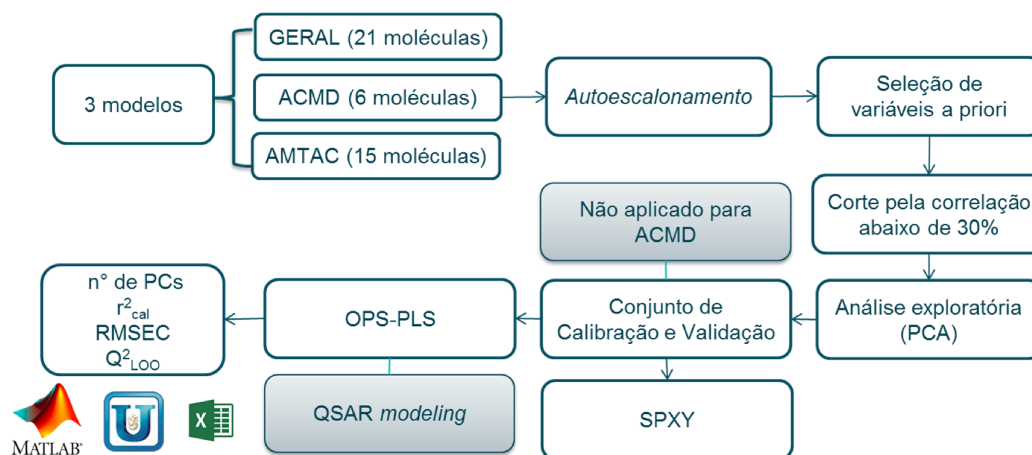
Onde o índice LV OPS significa a quantidade de variáveis latentes utilizadas pelo algoritmo OPS-PLS e LV_model significa a quantidade de variáveis latentes utilizadas no modelo.

Os melhores modelos foram escolhidos com base nos critérios de PCs, r^2_{cal} , RMSEC, e Q^2_{Loo} .

Visando a aplicação do modelo proposto no planejamento novos candidatos a fármacos, foi realizada a predição de novas moléculas com o intuito de prever a atividade biológica de moléculas que possam ser candidatos a fármacos.

A análise quimiométrica foi realizada usando os softwares The Unscrambler 9.7 da CAMO Process AS e o programa computacional Matlab (R2015b). O algoritmo SPXY foi utilizado por meio do pacote DATA_HAND_GUI (GOMES, 2012) em ambiente matlab. Um fluxograma dos procedimentos realizados é apresentado na Figura 5.

Figura 5 - Fluxograma da pesquisa



Fonte: Elaborada pelo autor, 2022.

5 RESULTADOS E DISCUSSÕES

5.1 Modelo biológico

Os resultados da viabilidade celular dos grupos de moléculas selecionadas (Pinheiro Segundo, 2020; Gouveia, 2017) dos grupos de moléculas ACMD (6 moléculas) e AMTAC (15 moléculas) podem ser vistas na tabela 1.

Tabela 1 – Resultado *in vitro* de viabilidade celular das moléculas dos grupos AMCD e AMTAC para atividade anticâncer com dosagem de 50 μ M contra linhagens de células humanas (HCT-116) para câncer de cólon.

Molécula	Potencial citotóxico (%)	Molécula	Potencial citotóxico (%)
ACMD-01	71.9	AMTAC-11	55.1
ACMD-02	61.9	AMTAC-13	51.8
ACMD-03	48.4	AMTAC-15	28.5
ACMD-04	21.7	AMTAC-16	46.6
ACMD-05	5.9	AMTAC-17	70.16
ACMD-06	19.8	AMTAC-18	54.3
AMTAC-02	58.66	AMTAC-19	88.56
AMTAC-04	82.95	AMTAC-21	70.6
AMTAC-06	93.8	AMTAC-22	67.2
AMTAC-09	49.4	AMTAC-23	33.41
AMTAC-10	62.2		

Fonte: Elaborada pelo autor, 2022.

As moléculas que apresentam acima de 70% de potencial citotóxico possuem muita atividade de inibição (PINHEIRO SEGUNDO, 2020), como as moléculas AMCD-01, AMTAC-04, AMTAC-06, AMTAC-17, AMTAC-19 e AMTAC-21.

5.2 Seleção de variáveis a priori

Em seguida, foram calculados os descritores moleculares para os três modelos: ACMD, AMTAC e GERAL. A partir do cálculo dos descritores, foi realizado a seleção de variáveis a priori com o objetivo de retirar variáveis que não apresentassem

informação química significativa. Posteriormente, foi realizado o corte pela correlação com significância abaixo de 0.3 (30%). A síntese desses resultados podem ser vistos na tabela 2.

Tabela 2 – Resultados da seleção de variáveis a priori para os modelos ACMD, AMTAC e GERAL

Modelos OPS	Matriz original	Seleção a priori	Corte (0.3)
ACMD	X (7 x 1444)*	X (7 x 1018)	X (7 x 547)
AMTAC	X (15 x 1444)	X (15 x 1067)	X (15 x 512)
GERAL	X (21 x 1444)	X (21 x 1073)	X (21 x 237)

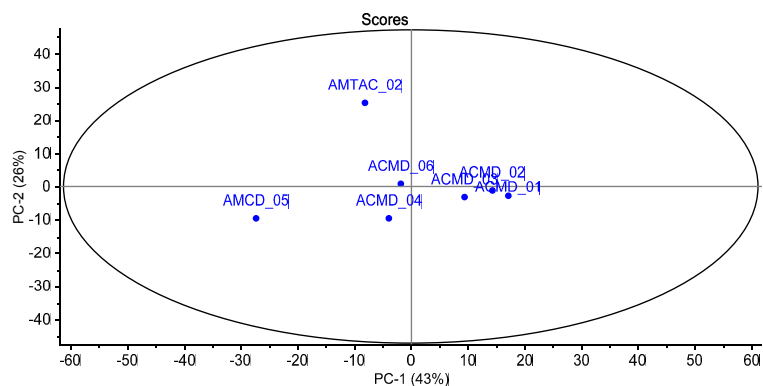
*Matrix **X** (7 amostras x 1444 variáveis)

Fonte: Elaborada pelo autor, 2022.

5.2.1 Análise exploratória

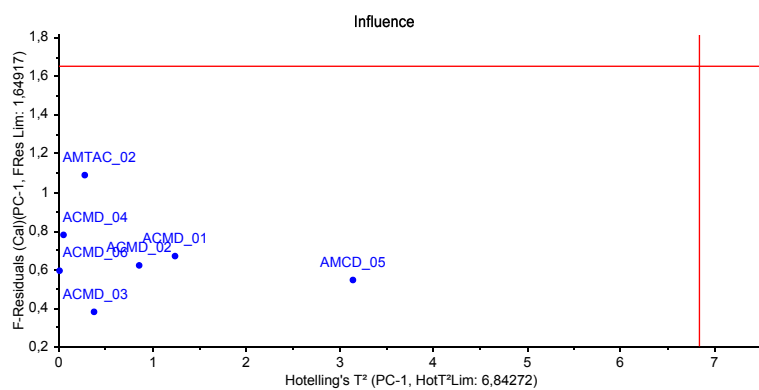
A análise PCA realizada a partir do corte pela correlação para os três modelos (ACMD, AMTAC e GERAL) possibilitou analisar o comportamento das amostras frente aos conjunto de variáveis. Os gráficos de escores podem ser vistos na figura 5. De acordo com o gráfico de escores, nota-se que na figura (5a) mostra um agrupamento entre as amostras ACMD 01, 02 e 03. A amostra AMTAC-02 está distante das demais, que é um sugestivo que existem diferenças entre as demais amostras. Na figura (5c) observa-se dois agrupamentos principais de amostras e duas amostras (AMTAC-15 e AMTAC-23) distantes das demais. Essas duas amostras estão distante entre si. Na figura (5e) nota-se a formação de dois agrupamentos, o primeiro formado apenas de amostras do grupo AMTAC, o segundo com amostras do grupo ACMD. Neste sentido, a partir da figura (5e), pode-se sugerir que os grupos ACMD e AMTAC apresentam diferenças significativas e não se misturam. Nenhuma amostra apresentou comportamento anômalo e nenhum valor elevado no gráfico de acordo com o gráfico de Influência.

Figura 6 - Gráficos de Escores da PCA do grupo ACMD



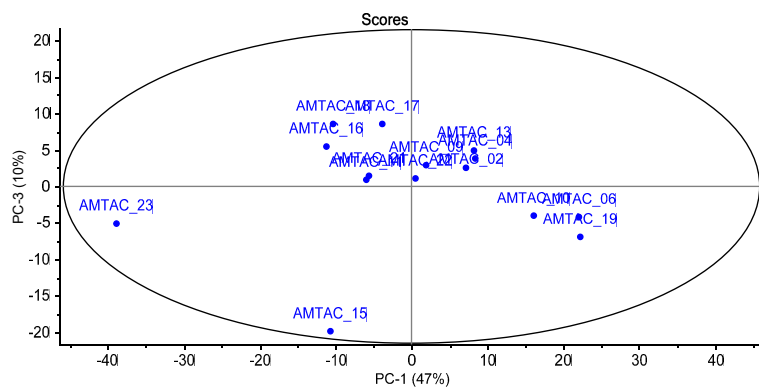
Fonte: Elaborada pelo autor, 2022.

Figura 7 - Gráficos de Influência da PCA do grupo ACMD



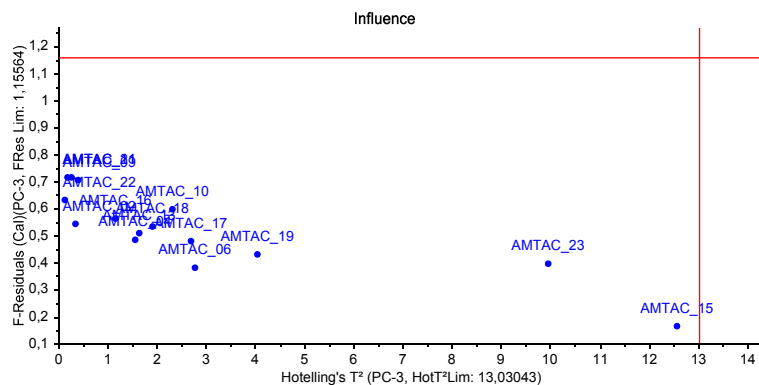
Fonte: Elaborada pelo autor, 2022.

Figura 8 - Gráficos de escores da PCA do grupo AMTAC



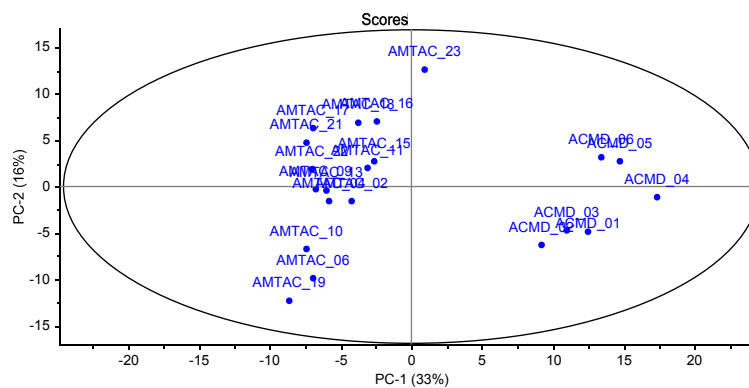
Fonte: Elaborada pelo autor, 2022.

Figura 9 - Gráficos de Influência da PCA do grupo AMTAC



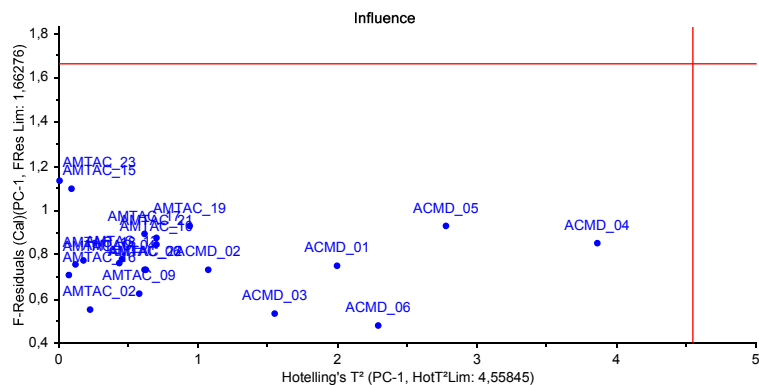
Fonte: Elaborada pelo autor, 2022.

Figura 10 - Gráficos de escores da PCA do grupo GERAL



Fonte: Elaborada pelo autor, 2022.

Figura 11 - Gráficos de Influência da PCA do grupo GERAL



Fonte: Elaborada pelo autor, 2022.

5.3 Análise dos modelos OPS-PLS

O conjunto de calibração e predição dos modelos foram construídos utilizando o algoritmo de seleção de amostras SPXY. Para o grupo ACMD foram utilizadas 6 moléculas do grupo ACMD para calibração e uma amostra do grupo AMTAC (AMTAC-02) para predição do modelo. Os grupos AMTAC e GERAL utilizaram 12 e 18 amostras para o conjunto de calibração, respectivamente, e 3 amostras para o conjunto de predição.

Os modelos OPS-PLS foram criados a partir das submatrizes obtidos a partir do corte pela correlação. Visando a redução do número de descritores, sucessivas seleções de variáveis foram realizadas a partir do modelo antecessor até resultar no menor número de variáveis significativas e que obtivesse uma boa capacidade preditiva. Para os modelos ACMD, AMTAC e GERAL foram construídos quatro, três e três submodelos OPS-PLS, respectivamente. Na tabela 3 pode ser visto a síntese dos resultados obtidos.

Tabela 3 – Seleção de variáveis OPS-PLS para os modelos ACMD, AMTAC e GERAL

Modelos OPS	Corte (0.3)	1° Seleção	2° Seleção	3° Seleção
ACMD	X (7 x 547)*	X (7 x 127)	X (7 x 105)	X (7 x 56)
AMTAC	X (15 x 512)	X (15 x 28)	X (15 x 7)	
GERAL	X (21 x 237)	X (21 x 58)	X (21 x 25)	

*Matrix **X** (7 amostras x 547 variáveis)

Fonte: Elaborada pelo autor, 2022.

A escolha dos melhores modelos ocorreu seguindo o princípio da parcimônia (“deve-se usar sempre o modelo mais simples possível”). Sendo assim, os melhores modelos foram: ACMD (**X** (7 x 56)), AMTAC (**X** (15 x 7) e GERAL (**X**(21 x 25). Suas respectivas figuras de mérito para o conjunto de calibração serão apresentadas de acordo com a tabela 4.

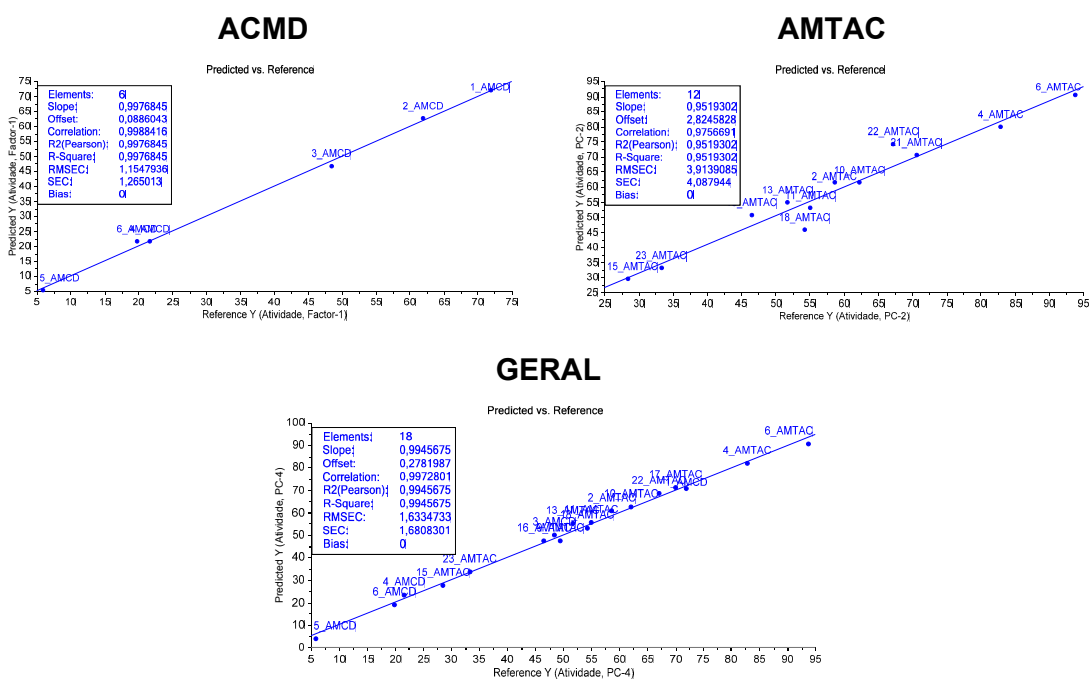
Tabela 4 – Figuras de mérito para o conjunto de calibração de regressão PLS para os modelos ACMD, AMTAC e GERAL

	ACMD	AMTAC	GERAL
PCs	1	2	4
RMSEC	1.15	3.91	1.63
r^2_{cal}	0.9976	0.9519	0.9945
Q^2_{Loo}	0.9976	0.8729	0.9945

Fonte: Elaborada pelo autor, 2022.

De acordo com as figuras de mérito, percebe-se que os três modelos apresentam valores de RMSEC relativamente baixos, baseados na porcentagem da variação da atividade biológica. A figura 6 mostra os gráficos dos valores medidos vs valores previstos pelos modelos ACMD, AMTAC e GERAL para o conjunto de calibração. Observa-se uma dispersão homogênea das amostras em torno da bissetriz, o que mostra uma boa performance para os modelos.

Figura 6 – Gráficos dos valores medidos vs valores previsto pelos modelos ACMD, AMTAC e GERAL no conjunto de calibração



Fonte: Elaborada pelo autor, 2022.

5.4 Validação dos modelos

Os valores obtidos para o conjunto de validação (% atividade biológica prevista e medida) para os três modelos e seus respectivos desvios padrões podem ser visualizados na tabela na tabela 5. As figuras de mérito para o conjunto de validação pode ser visto na tabela 6.

Tabela 5 – Valores medidos vs preditos e desvios padrões para os modelos AMCD, AMTAC e GERAL

Modelos		Atividade biológica medida (%)	Atividade biológica prevista (%)	Desvio padrão
ACMD	AMTAC-02	58.66	57.75	6.68
AMTAC	AMTAC-09	49.40	50.63	3.49
	AMTAC-17	70.16	71.82	4.37
	AMTAC-19	88.56	89.93	4.68
GERAL	ACMD-02	61.90	58.05	4.84
	AMTAC-19	88.56	94.78	4.02
	AMTAC-21	70.60	69.80	5.86

Fonte: Elaborada pelo autor, 2022.

Em relação ao valores preditos, observa-se que ambos os modelos apresentaram ótima capacidade preditiva. Em relação aos desvios padrões, o grupo ACMD, obteve o maior desvio, com 6.68 para amostra AMTAC-02. É importante ressaltar que mesmo a molécula sendo de outro grupo AMTAC, a atividade predita foi muito parecida com a atividade medida. Para o grupo AMTAC, ambas as amostras apresentaram desvio menor que 5 e valores preditos próximos das atividade medidas, o que corrobora sua ótima capacidade preditiva. Para o grupo GERAL, as amostras ACMD-02 e AMTAC-19 apresentaram os menores desvios entre os grupos, com 4.84 e 4.02, respectivamente. Os três grupos apresentaram comportamento robusto, uma vez que seus parâmetros quase não sofreram alterações em relação as condições distintas para cada modelos. Não foi possível obter o valor do r^2_{pred} do modelo ACMD por causa do número de amostras utilizadas na validação.

Tabela 6 – Figuras de mérito para o conjunto de validação da seleção de variáveis OPS-PLS para os modelos ACMD, AMTAC e GERAL

	ACMD	AMTAC	GERAL
PCs	1	2	4
r^2_{pred}		0.9919	0.8533
RMSEP	0.91	1.43	4.25

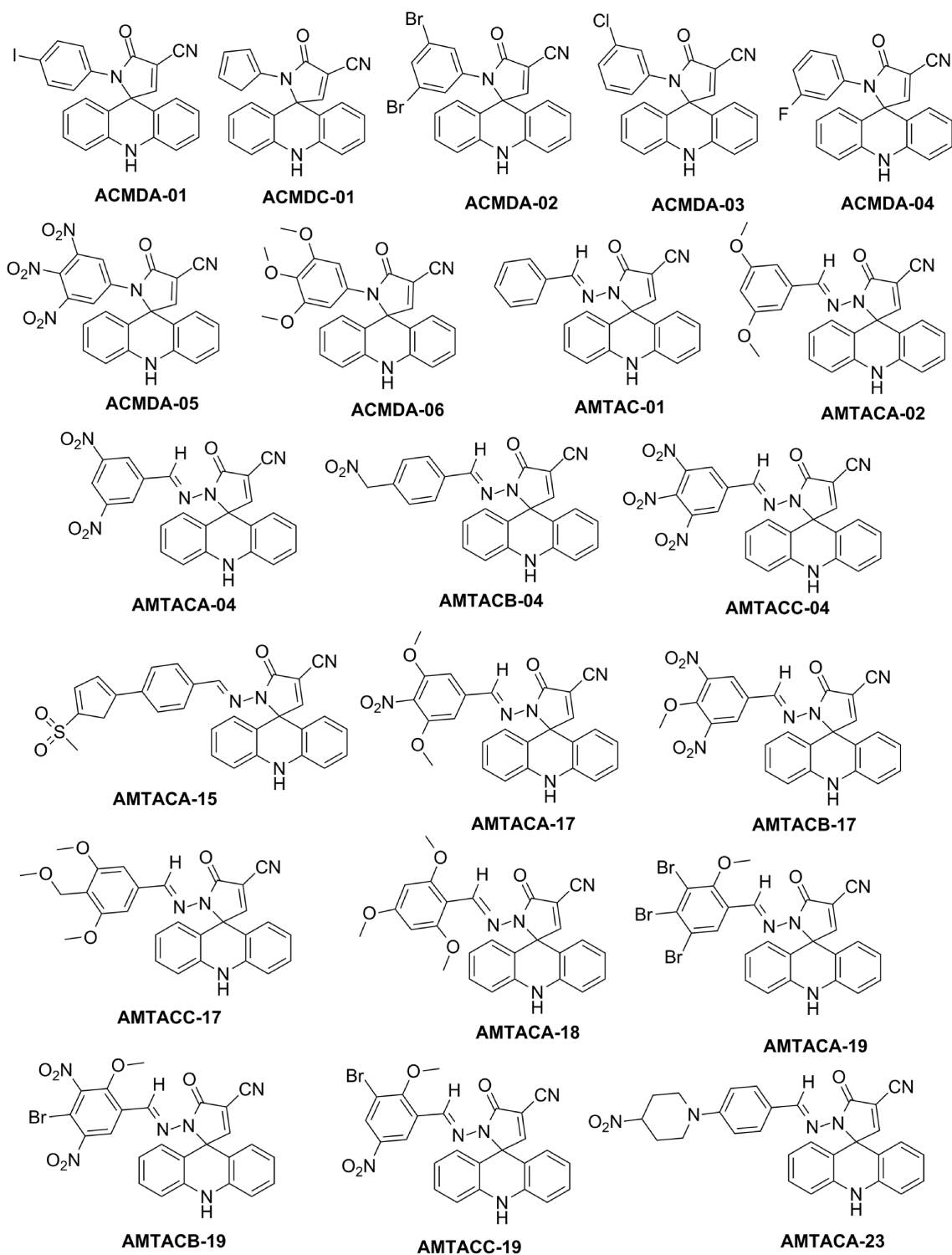
Fonte: Elaborada pelo autor, 2022.

Em relação as figuras de mérito dos modelos AMCD, AMTAC e GERAL, observa-se que o aumento do RMSEP é diretamente proporcional ao aumento do número de amostras. Isto está relacionado com o aumento do erro quadrático da previsão da qualidade da atividade medida. Além disso, o número de PCs fortalece esse indicativo, visto que os modelos necessitam de mais PCs para poderem descrever a máxima variância dos dados.

5.5 Predição de novas moléculas como possíveis candidatos a fármacos

Foram desenvolvidas 20 moléculas (7 ACMD e 13 AMTAC) utilizando a Homologação de estrutura como estratégia de desenvolvimento de fármacos para a previsão da atividade biológica dos candidatos a fármacos a partir de alterações de grupos funcionais das moléculas utilizadas no conjunto de calibração. Além dessas, foi inseiro o composto AMTAC-01 foi previamente sintetizado por Gouveia (2017). Entretanto, o ensaio para avaliar a atividade biológica ainda não foi realizado. As novas estruturas podem ser visualizadas na tabela 7.

Figura 7 – Estruturas químicas das novas moléculas de predição



Fonte: Elaborada pelo autor, 2022.

As novas moléculas passaram pelo mesmo procedimento de tratamento de dados realizado no conjunto de calibração. Em seguida, suas atividades biológicas foram previstos utilizando os melhores modelos para ACMD, AMTAC e GERAL. Os respectivos valores previstos e os desvios padrões podem ser vistos na tabela 7, 8 e 9.

Tabela 7 – Valores da atividade biológica prevista das moléculas do modelo ACMD

Moléculas	Atividade biológica prevista (%)	Desvio padrão
ACMDA-01	61.003	5.053
ACMDA-01	52.528	6.099
ACMDA-02	55.016	4.962
ACMDA-03	49.754	4.06
ACMDA-04	40.877	4.761
ACMDA-05	0.356	8.361
ACMDA-06	27.81	5.803

Fonte: Elaborada pelo autor, 2022.

Os valores da atividade prevista para o modelo ACMD apresentaram um valor predito entre 0.3 a 61% e o desvio padrão entre 4.06 a 8.3. Moléculas com atividade biológica superior a 70% possuem alto potencial de inibição. Não foi verificado nenhuma molécula com boa atividade.

Tabela 8 – Valores da atividade biológica prevista das moléculas do modelo AMTAC

Moléculas	Atividade biológica prevista (%)	Desvio padrão
AMTAC-01	58.462	5.991
AMTACA-02	70.543	3.639
AMTACA-04	68.595	3.702
AMTACB-04	73.441	3.892
AMTACC-04	42.97	3.585
AMTACA-15	37.415	10.916
AMTACA-17	50.586	3.187
AMTACB-17	69.301	2.954
AMTACC-17	62.137	5.009
AMTACA-18	41.695	4.747
AMTACA-19	66.704	6.979
AMTACB-19	61.158	5.507
AMTACC-19	66.485	4.235
AMTACA-23	42.433	10.409

Fonte: Elaborada pelo autor, 2022.

Os valores da atividade biológica prevista para o modelo AMTAC apresentaram resultados satisfatórios. A atividade predita das moléculas variou entre 37.41 e 73.44% e o desvio padrão entre 3.18 e 10.91. Desse modo, destacam-se as moléculas AMTACA-02 (70.54%) e AMTACB-04 (73.44%). Neste sentido, ao todo foram identificados 2 moléculas com atividade biológica com alta atividade de inibição.

Tabela 9 – Valores previstos das moléculas do modelo GERAL

Moléculas	Atividade biológica predita (%)	Desvio padrão
ACMDA-01	62.608	6.096
ACMDA-01	68.445	6.005
ACMDA-02	44.036	7.677
ACMDA-03	43.895	5.039
ACMDA-04	26.783	4.896
ACMDA-05	-22.082	12.698
ACMDA-06	42.869	6.783
AMTA-01	84.466	5.106
AMTACA-02	54.453	4.988
AMTACA-04	47.454	7.442
AMTACB-04	76.212	5.749
AMTACC-04	47.104	7.896
AMTACA-15	22.048	9.146
AMTACA-17	63.825	4.966
AMTACB-17	66.201	4.83
AMTACC-17	61.696	5.719
AMTACA-18	51.702	7.137
AMTACA-19	47.766	4.943
AMTACB-19	46.338	8.19
AMTACC-19	42.076	5.745
AMTACA-23	27.14	3.338

Fonte: Elaborada pelo autor, 2022.

Os valores previstos para o modelo GERAL apresentaram resultados complementares para os modelos ACMD e AMTAC. Para as amostras do grupo ACMD, os valores previstos variaram entre -22.08 e 68.44% e desvios padrões entre 4.89 e 12.69. É interessante ressaltar que a adição dos grupos nitro reduziu a nível negativo a atividade biológica da molécula ACMDA-05. Em comparação aos modelos ACMD e AMTAC, os resultados previstos foram similares, exceto pelas moléculas ACMDA-04, ACMDA-05 e ACMDA-06. As amostras do grupo AMTAC apresentaram valores previstos entre 27.14 e 84.46%, com desvios padrões entre 4.83 e 9.14. Os compostos AMTAC-1 (84.46%) e AMTACB-04 (76.21%) foram considerados com alta atividade de inibição.

6 CONSIDERAÇÕES FINAIS

O modelo desenvolvido utilizando QSAR-2D e derivados espiro-acridínicos com atividade anticâncer de cólon se mostrou viável para o avaliar novos candidatos a fármacos com alta atividade de inibição para o câncer de cólon. O modelo ACMD apresentou bons resultados, mas seu número de amostras ainda é um fator limitante. O modelo AMTAC mostrou-se promissor e preveu 2 novos compostos com uma ótima atividade de inibição. O modelo GERAL serviu como comparativo entre os modelos ACMD e AMTAC. Para continuação do estudo, é necessário avaliar quais são os descritores moleculares mais significantes para o aumento da atividade biológica em cada modelo. Assim, permitindo o avanço direcionado a proposta de novas moléculas e descoberta do seu possível mecanismo de ação. Os resultados apresentados possuem grande relevância, visto que testes de validação de amostras reais demonstraram que os modelos QSAR foram confiáveis e preditivos. Neste sentido, os modelos podem ser utilizados por pesquisadores que desejam sintetizar e avaliar novos compostos semelhantes às estruturas dos derivados espiro-acridinínicos para o planejamento de novas moléculas com alta atividade contra o câncer de cólon.

7 PERSPECTIVAS

- Avaliar a significância dos descritores moleculares que mais contribuem para atividade biológica de cada modelo, favorecendo a descoberta de grupos funcionais chaves para alta atividade dos compostos, como seu mecanismo de ação.
- Devido a limitação de moléculas do grupo ACMD, é necessário a síntese de mais moléculas para avaliar a atividade biológica contra o câncer de cólon de uma forma mais efetiva.
- Sintetizar a molécula AMTACB-04 e avaliar biologicamente a molécula AMTAC-01 que obtiveram alta atividade biológica prevista, o que servirá como uma validação externa do modelo AMTAC desenvolvido.
- Abranger o estudo utilizando descritores moleculares 3D, como os geométricos, afim de ampliar as informações que até então são um fator limitante para os modelos desenvolvidos.
- Desenvolver outros métodos de QSAR, como o 3D, com o objetivo de utilizar descritores de campo em uma caixa 3D virtual, possibilitando o cálculo das interações eletrostáticas e de Van der Waals entre os átomos das moléculas do conjunto de dados.
- Avaliar o uso de outros métodos quimiométricos combinado a seleção de variáveis, como os algoritmos heurísticos, visando a descoberta de métodos mais precisos e robustos.

REFERÊNCIAS

DE ALMEIDA, S. M.; LAFAYETTE, E. A.; SILVA, W. L.; LIMA SERAFIM, V.; MENEZES, T. M.; NEVES, J. L.; RUIZ, A. L.; CARVALHO, J. E.; MOURA, R. O.; BELTRÃO, E. I.; CARVALHO JÚNIOR, L. B.; LIMA, M. D. New spiro-acridines: DNA interaction, antiproliferative activity and inhibition of human DNA topoisomerases. **Int J Biol Macromol**, [s. l.], v. 16, n. 92, p. 457-475. 2016.

BYSTRZANOWSKA, M.; TOBISZEWSKI, M. Chemometrics for selection, prediction, and classification of sustainable solutions for green chemistry—a review. **Symmetry**, [s. l.], v. 12, n. 12, p. 1–21. 2020.

CHIAPPINI, A. F.; ALLEGRINI, F.; GOICOECHEA, C. H.; OLIVIERI, C. A. Sensitivity for Multivariate Calibration Based on Multilayer Perceptron Artificial Neural Networks. **Analytical Chemistry**, [s. l.], v. 92, n. 18, p. 12265–12272. 2020.

COOK, R. D.; FORZANI, L. Envelopes: A new chapter in partial least squares regression. **Journal of Chemometrics**, [s. l.], v. 34, n. 10, p. 1–20. 2020.

DE PAULO, H. E.; FOLLI, S. G.; NASCIMENTO, C. H. M.; MORO, K. M.; DA CUNHA, P. H. P.; CASTRO, R. V. E.; NETO, C. A.; FILGUEIRAS, R. P. Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. **Fuel**, [s. l.], v. 279, n. May, p. 118462. 2020.

DUARTE, S. S.; SILVA, D. K. F.; LISBOA, T. M. H.; GOUVEIA, R. G.; FERREIRA, R. C.; DE MOURA, R. O.; DA SILVA, J. M.; DE ALMEIDA, L. É.; RODRIGUES MASCARENHAS, S.; DA SILVA, P. M.; FARIAS, D. F.; DA COSTA, R. S. J. A.; DE PAULA, M. K. C.; GONÇALVES, J. C. R.; SOBRAL, M. V. Anticancer effect of a spiro-acridine compound involves immunomodulatory and anti-angiogenic actions. **Anticancer Research**, [s. l.], v. 40, n. 9, p. 5049–5057. 2020.

DUARTE, S. S.; SILVA, D. K. F.; LISBOA, T. M. H.; GOUVEIA, R. G.; DE ANDRADE, C. C. N.; DE SOUSA, V. M.; FERREIRA, R. C.; DE MOURA, R. O.; GOMES, J. N. S.; DA SILVA, P. M.; DE LOURDES, A. A. A. F.; KEESEN, T. S. L.; GONÇALVES, J. C. R.; BATISTA, L. M.; SOBRAL, M. V. Apoptotic and antioxidant effects in HCT-116 colorectal carcinoma cells by a spiro-acridine compound, AMTAC-06. **Pharmacol Rep**, [s. l.], v. 74, n. 3, p. 545-554. 2022.

FERREIRA, M. M. C. **Quimiometria: Conceitos, Métodos e Aplicações**. Campinas: Editora da Unicamp, 2015, 493p.

GOMES, A. A. **Algoritmo das projeções sucessivas à seleção de variáveis em regressão PLS**. 2012. 121f. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa. 2012.

GOUVEIA, G. R. **Síntese, caracterização estrutural e avaliação dos possíveis mecanismos de ação dos derivados espiro-acridínicos**. 2017. Dissertação (Mestrado em Ciências Farmacêuticas) – Universidade Estadual da Paraíba, Paraíba, 2017.

INCA. Estatísticas de câncer. INCA, 2020. Disponível em:

<https://www.inca.gov.br/numeros-de-cancer>. Acessado em: 30/04/2022.

HALDER, A. K.; DIAS SOEIRO CORDEIRO, M. N. Advanced in Silico Methods for the Development of Anti-Leishmaniasis and Anti-Trypanosomiasis Agents. **Current Medicinal Chemistry**, [s. l.], v. 27, n. 5, p. 697–718. 2020.

JIMÉNEZ-LUNA, J.; GRISONI, F.; WESKAMPB, N.; SCHNEIDER, G. Artificial intelligence in drug discovery: recent advances and future perspectives. **Expert Opinion on Drug Discovery**, [s. l.], v. 16, n. 9, p. 949–959. 2021.

LIN, X.; LI, X.; LIN, X. A review on applications of computational methods in drug screening and design. **Molecules**, [s. l.], v. 25, n. 6, p. 1–17. 2020.

MARTINS, J. P. A.; FERREIRA, M. M. C. QSAR modeling: um novo pacote computacional open source para gerar e validar modelos QSAR. **Química Nova**, [s. l.], v. 36, n. 4, p. 554–560. 2013.

MEHMOOD, T.; SÆBØ, S.; LILAND, K. H. Comparison of variable selection methods in partial least squares regression. **Journal of Chemometrics**, [s. l.], v. 34, n. 6, p. 1–14. 2020.

PINHEIRO SEGUNDO, S. A. M. **Desenvolvimento, avaliação preliminar da atividade antiproliferativa e incremento de solubilidade de novos derivados espiroacridínicos**. 2020. Tese (Doutorado em Ciências Farmacêuticas) – Universidade Federal de Pernambuco, Recife. 2020.

MURATOV, E. N.; BAJORATH, J.; SHERIDAN, R.P.; TETKO, I.V.; FILIMONOV, D.; POROIKOV, V.; OPREA, T. I.; BASKIN, I.I.; VARNEK, A.; ROITBERG, A.; ISAYEV, O.; CURTAROLO, S.; FOURCHES, D.; COHEN, Y.; ASPURU-GUZIK, A.; WINKLER, D.A.; AGRAFIOTIS, D.; CHERKASOV, A.; TROPSHA, A. QSAR without borders. **Chemical Society Reviews**, [s. l.], v. 49, n. 11, p. 3525–3564. 2020.

OLIVERI, P.; MALEGORI, C.; CASALE, M. **Chemometrics: multivariate analysis of chemical data**. Second Editioned. [S. l.]: Elsevier Inc. 2020.

OYEDELE, A. S.; BOGAN, D. N.; OKORO, C. O. Synthesis, biological evaluation and virtual screening of some acridone derivatives as potential anticancer agents. **Bioorg Med Chem**. [s. l.], v. 28, n. 9, p. 115426. 2020.

ROQUE, V. J.; CARDOSO, W.; PETERNELLI, A. L.; TEÓFILO, F. R. Comprehensive new approaches for variable selection using ordered predictors selection. **Analytica Chimica Acta**, [s. l.], v. 1075, p. 57–70. 2019.

ROY, K.; KAR, S.; DAS, R. N. **A Primer on QSAR/QSPR Modeling: Fundamental Concepts** (Springer Briefs in Molecular Sciences). [S. l.: s. n.]. 2015.

SHARMA, V.; GUPTA, M.; KUMAR, P.; SHARMA, A. A Comprehensive Review on Fused Heterocyclic as DNA Intercalators: Promising Anticancer Agents. **Current Pharmaceutical Design**, [s. l.], v. 27, n. 1, p. 15–42. 2020.

SINGH, R. P.; AZIZ, M. N.; GOUT, D.; FAYAD, W.; EL-MANAWATY, M. A.; LOVELY, C. J. Novel thiazolidines: Synthesis, antiproliferative properties and 2D-QSAR studies. **Bioorganic and Medicinal Chemistry**, [s. l.], v. 27, n. 20, p. 115047. 2019.

TEÓFILO, R. F.; MARTINS, J. P. A.; FERREIRA, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. **Journal of Chemometrics**, [s. l.], v. 23, n. 1, p. 32–48. 2009.

WANG, L.; DING, J.; PAN, L.; CAO, D.; JIANG, H.; DING, X. Quantum chemical descriptors in quantitative structure–activity relationship models and their applications. **Chemometrics and Intelligent Laboratory Systems**, [s. l.], v. 217, n. July, p. 104384. 2021.

WHO. Global health estimates 2020: deaths by cause, age, sex, by country and by region 2000-2019. WHO, 2020. Disponível em: <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/gheleading-causes-of-death>. Acessado em: 30/04/2022.

