Ministério da Educação
Universidade Federal de Goiás
Faculdade de Farmácia
Programa de Pós-Graduação em Ciências Farmacêuticas

FLÁVIA CRISTINA DA SILVA

Desenvolvimento de modelos de QSAR para identificação de substratos e inibidores de CYP3A4

Goiânia
2015

FLÁVIA CRISTINA DA SILVA

Desenvolvimento de modelos de QSAR para identificação de substratos e inibidores de CYP3A4

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas da Faculdade de Farmácia da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciências Farmacêuticas.

Área de Concentração: Fármacos e Medicamentos

Orientadora: Profa. Dra. Carolina Horta Andrade

Goiânia
2015

Ficha catalográfica elaborada automaticamente
com os dados fornecidos pelo(a) autor(a), sob orientação do Sibi/UFG.

Cristina da Silva, Flávia
Desenvolvimento de modelos de QSAR para identificação de substratos e inibidores de CYP3A4 [manuscrito] / Flávia Cristina da Silva. - 2015.
XVII, 90 f.

Orientador: Profa. Dra. Carolina Horta Andrade.
Dissertação (Mestrado) - Universidade Federal de Goiás, Faculdade Farmácia (FF) , Programa de Pós-Graduação em Ciências Farmacêuticas, Goiânia, 2015.
Bibliografia. Apêndice.
Inclui siglas, abreviaturas, tabelas, lista de figuras, lista de tabelas.

1. QSAR. 2. CYP3A4. 3. metabolismo de fármacos . 4. substrato . 5. inibidor . I. Horta Andrade, Carolina , orient. II. Título.

Folha de Aprovação

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciências Farmacêuticas da Universidade Federal de Goiás, em 26 de fevereiro de 2015, pela mestranda Flávia Cristina da Silva.

Banca Examinadora:

Profa. Dra. Carolina Horta Andrade (FF/UFG)
Presidente

Prof. Dr. Vinícius Barreto da Silva (PUC/GO)

Profa. Dra. Valéria de Oliveira (FF/UFG)

DEDICATÓRIA

Aos meus pais Valdeci e Aurinete e a minha avó Maria eu dedico este trabalho, por todo apoio e carinho que foi me dado durante toda essa jornada de trabalho, sempre me motivando a realizar meus sonhos e me ensinando o verdadeiro sentido do amor e cuidado.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me permitir a existência, por direcionar minha vida e por permitir que mais essa etapa na minha vida fosse cumprida, tendo a certeza que sempre esteve ao meu lado em todos os momentos.

À profa. Carolina Horta Andrade, pela paciência, atenção, confiança e orientação, por ter me aberto às portas á pesquisa mesmo diante de minhas dificuldades, para que pudéssemos concretizar este trabalho, meus sinceros agradecimentos.

Ao MSc. Rodolpho de Campos Braga, por toda orientação, suporte, contribuição e discussões científicas que me engrandeceram muito durante essa caminhada.

Ao Grupo de Pesquisa do Laboratório de Modelagem Molecular – LabMol, que me acompanhou durante esses dois anos e por toda amizade construída, sempre estarão na minha memória e coração, em especial ao Meryck por toda ajuda durante essa caminhada.

Aos meus pais, Valdeci e Aurinete pelo carinho, incentivo e apoio em toda minha existência me ensinando a importância da educação, dedicação e trabalho.

A minha irmã, Fernanda por todo amor e torcida em importantes momentos da minha vida acadêmica.

Um agradecimento especial à minha avó, Maria das Graças por compartilhar minhas alegrias e por ter sempre me apoiado, por ter acreditado em mim e pela oração. E ao Kiko por apenas ter existido em minha vida.

As minhas grandes amigas, Karen, Benaia e Géssyka pelo apoio e por sempre acreditar em mim.

À todos os professores e funcionários da Faculdade de Farmácia da UFG.

À CAPES, FAPEG e CNPq, pelo apoio financeiro.

À todas as pessoas que, direta ou indiretamente, contribuíram para a execução dessa Dissertação de Mestrado.

EPÍGRAFE

"Para achar água é preciso descer terra adentro Encharcar-se no lodo.
Mas há os que preferem olhar os céus, e esperar pelas chuvas"

Osvaldo Viana Filho

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Acc	Acurácia
AM	Aprendizado de máquina
ADME/Tox	Absorção, distribuição, metabolismo, excreção e toxicidade
2D	Bidimensional ou segunda dimensão
3D	Tridimensional ou terceira dimensão
CYP450	Citocromo P450
CADD	Computer-Aided Drug Design
CAS	Chemical Abastract Services
CoMFA	*Comparative Molecular Fields Analysis*
CoMSIA	*Comparative Molecular Similarity Index Analysis*
DA	Domínio de aplicabilidade
DM	Dinâmina molecular
DDIs	Drug-drug interactions
ES	Especificidade
FDA	*Food and Drug Administration*
IUPAC	*International Union of Pure and Applied Chemistry*
IND	*Invetigational New Drug*
IA	Inteligência artificial
k-NN	Método do vizinho mais proximo
KSAR	Protocolo *in house* para o preparo e padronização do conjunto de dados
LBDD	*Ligand-based drug design*
MPPs	Mapas de probabilidade predita
MAS	Molecular Shape Analysis
NCE	*New chemical entities*
NADPH	Nicotinamida adenina dinucleotídeo fostato
OECD	*Organization for Economic Co-operation and Development*
P&D	Pesquisa e desenvolvimento
PDB	*Protein Data Bank*
PLSDA	*Partial least squares discriminant analysis*
QSAR	*Quantitative-structure-activity relationship*
qHTS	*Quantitative High-throughput screening*

RESUMO

A descoberta e o desenvolvimento de fármacos consistem um processo complexo, sendo necessária a integração de várias áreas estratégicas como conhecimento, inovação, tecnologia, gerenciamento e altos investimentos em Pesquisa, Desenvolvimento e Inovação (PD&I). Nenhum fármaco pode ser aprovado para uso em humanos sem que antes passe por extensivos estudos que visem garantir sua eficácia e segurança. Um fármaco que inibe a atividade metabólica de uma enzima da família citocromo P450 (CYP450), pode afetar a farmacocinética de outros fármacos, resultando em interações fármaco-fármaco (DDIs), que podem conduzir potencialmente a efeitos colaterais e tóxicos. As principais enzimas oxidativas responsáveis pelo metabolismo de fármacos possuem como principais representantes a superfamília CYP450, em que a isoforma CYP3A4 é a mais importante, pois é responsável por metabolizar aproximadamente 50 % dos fármacos disponíveis no mercado. Diversos métodos computacionais têm sido desenvolvidos como estratégia para predizer o metabolismo humano nos primeiros estágios de pesquisa e desenvolvimento de fármacos. Modelos *in silico* do metabolismo apresentam vantagens como maior rapidez, menor custo e maior facilidade de operação, quando comparados aos modelos tradicionais *in vitro* e *in vivo*. O trabalho teve como objetivo central o desenvolvimento de modelos de Relações Quantitativas entre estrutura química e atividade/propriedade (QSAR/QSPR) robustos e preditivos, visando identificar substratos e inibidores de CYP3A4. Para isso, foram compilados, integrados e preparados os maiores conjuntos de dados disponíveis na literatura de substratos e inibidores de CYP3A4. Vários modelos de QSAR foram gerados e validados para ambas as propriedades usando um fluxo de trabalho que contemplou criteriosamente as recomendações da *Organization for Economic Co-operation Development* (OECD). A combinação de diferentes descritores e métodos de aprendizado de máquina levaram a obtenção de modelos QSAR robustos e consistentes, com taxa de classificação correta (CCR) que variam entre 0,65-0,83 e cobertura de 0,69-0,89,demonstrando valores estatisticamente significativos para classificação com alta precisão de compostos em substratos ou não substratos de CYP3A4. O modelo Morgan-RF binário gerado para classificar compostos em inibidores e não inibidores se mostraram também altamente robusto e preditivo com valores de sensibilidade de 0,77 e acurácia de 0,76, e o modelo Morgan-RF multiclasse obteve valores de 0,68 para sensibilidade e 0,69 para acurácia. O mapa de probabilidade predita se mostrou útil, pois conseguiu codificar fragmentos estruturais importantes para classificar compostos em inibidores ou não inibidores de CYP3A4. Como conclusões foram desenvolvidos e validados diversos modelos de QSAR para prever a interação com a enzima CYP450 que podem ser úteis nos estágios iniciais do desenvolvimento de novos fármacos. O próximo passo será a disponibilização online dos modelos obtidos no servidor do LabMol (http://labmol.farmacia.ufg.br).

Palavras-chave: QSAR, *in silico*, metabolismo de fármacos, substrato, inibidor, CYP3A4.

ABSTRACT

The discovery and development of drugs consist of a complex process, requiring the integration of various strategic areas such as knowledge, innovation, technology, management and high investments in Research, Development and Innovation (RD&I). No drug can be approved for use in humans without first go through extensive studies aimed at ensuring its effectiveness and safety. On the other hand, a drug that inhibits the activity of a metabolic enzyme cytochrome P450 family (CYP450) can affect the pharmacokinetics of other drugs, resulting in drug-drug interactions (DDIs), which potentially lead to side effects and toxic effects. The main oxidative enzymes responsible for drug metabolism have as main representatives CYP450 superfamily, wherein the CYP3A4 isoform is the most important because it is responsible for metabolizing approximately 50% of the drugs on the market. Several computational methods have been developed as a strategy to predict human metabolism in the early stages of research and development of drugs. In silico models of metabolism have advantages such as faster, lower cost and ease of operation when compared to traditional models in vitro and in vivo. The work aimed mainly at the development of Quantitative Relations between models chemical structure and activity / property (QSAR / QSPR) robust and predictive, to identify CYP3A4 substrates and inhibitors. To this were collected, integrated and prepared larger data sets available in the literature substrates and inhibitors of CYP3A4. Several QSAR models were generated and validated for both properties using a workflow that contemplated carefully the recommendations of the Organization for Economic Co-operation Development (OECD). The combination of different descriptors and machine learning methods have led to obtain robust and predictive QSAR models, with correct classification rate (CCR) ranging from 0.65 to 0.83 and 0.69 to 0.89 of coverage, showing a statistically significant values for classification of compounds with high accuracy whether or not substrates of CYP3A4 substrates. The binary Morgan RF-generated model to classify compounds inhibitors and non-inhibitors also proved highly robust and predictive with sensitivity values of 0.77 and accuracy of 0.76, and the Morgan-RF model multiclass obtained values of 0.68 sensitivity and 0.69 for accuracy. The map of predicted probability proved useful as it could encode major structural fragments to classify compounds inhibitors or not CYP3A4 inhibitors. In conclusion, have been developed and validated many QSAR to predict the interaction with the CYP450 enzyme that may be useful in the early stages of the development of new drugs. The next step is the online availability of the models obtained in LabMol server (http://labmol.farmacia.ufg.br).

Keywords: QSAR, *in silico*, drug metabolism, substrate, inhibitor, CYP3A4.

1 INTRODUÇÃO

1.1 Metabolismo de fármacos

O metabolismo de fármacos é o processo desenvolvido pelo organismo, em que ocorrem modificações químicas na estrutura de fármacos, com o objetivo de promover sua remoção do organismo. Em alguns casos, o metabolismo conduz a inativação do fármaco. Em algumas vezes, as transformações metabólicas podem produzir metabólitos tóxicos, com implicações toxicológicas potencias. Dessa forma, o metabolismo de fármacos evita o acúmulo e, posteriormente, a intoxicação por um xenobiótico (KUMAR; SURAPANENI, 2001; MONTELLANO, 2010).

Nenhum fármaco pode ser aprovado para ser usado em humanos sem que antes passe por estudos que visem determinar sua eficácia e segurança. A elucidação do metabolismo de fármacos constitui uma etapa importante e necessária para essa avaliação (STROHMEIER et al., 2011).

O metabolismo de fármacos compreende o conjunto de reações enzimáticas que realizam biotransformações específicas na estrutura química do xenobiótico, resultando em compostos com polaridade crescente e com maior facilidade de eliminação pelo organismo, impedindo desta forma, que estes compostos fiquem por tempo indefinido no organismo. Em alguns casos, entretanto, o metabolismo pode conduzir a compostos com atividade biológica, produzindo compostos farmacologicamente mais ativos ou mesmo tóxicos, justificando a necessidade de estudos farmacológicos e toxicológicos mais detalhados dos metabólitos (SMITH; DALVIE, 2011; TESTA; PEDRETTI; VISTOLI, 2012).

Os fármacos, assim como outros xenobióticos (por exemplo, solventes industriais, pesticidas etc.), são metabolizados por distintos sistemas enzimáticos, visando assegurar seu processo de inativação e eliminação. Várias enzimas encontradas em humanos, sejam específicas ou não, catalisam o metabolismo de xenobióticos de forma estereoespecífica, com o objetivo de transformar fármacos lipofílicos em metabólitos mais hidrofílicos, favorecendo a eliminação por via renal (VILLAGRA et al., 2011). Embora o fígado seja o principal sítio de metabolismo de fármacos, esse processo também pode ocorrer em outros órgãos como, intestino, rins, pulmões e cérebro.

O metabolismo de fármacos ocorre em duas fases bem distintas e caracterizadas, denominadas de reações de Fase I (biotransformação) e Fase II (conjugação) (MONTELLANO, 2010).

1.1.1 Reações de Fase I

As reações de Fase I são reações de funcionalização que se caracterizam por envolver reações de oxidação, redução e hidrólise, as quais mais frequentemente resultam na obtenção de metabólitos mais hidroxilados. Nessas reações, um grupo funcional é introduzido na molécula do fármaco, ou um grupo funcional existente é modificado, ou ainda um grupo funcional ou um sítio receptor para reações de transferência de fase II é exposto, levando a metabólitos mais hidrofílicos, tornando assim o xenobiótico mais polar e consequentemente mais facilmente excretado (MONTELLANO, 2010).

As reações de oxidação de Fase I são as mais comuns no metabolismo de xenobióticos. Dentre as enzimas, destaca-se o citocromo P450 (CYP450), um sistema enzimático da hemeproteína oxidativa, que se apresenta solúvel no citoplasma de células procarióticas e localiza-se na membrana de mitocôndrias e retículo endoplasmático liso do fígado e de outros tecidos extra-hepáticos de células eucarióticas (JOHNSTON et al., 2008).

As reações metabólicas de redução promovem uma modificação na estrutura do substrato através da adição de hidrogênios em sistemas contendo ligações duplas. Essas reações podem ocorrer em nível microssomal por ação do complexo enzimático NADPH-citocromo P450 redutase. Essas enzimas são ligadas à membrana e frequentemente catalisam a redução de grupo nitro ($R\text{-}NO_2$) ao metabólito amina ($R\text{-}NH_2$) (TESTA; PEDRETTI; VISTOLI, 2012; LIMA, 2015).

As reações hidrolíticas são catalisadas por um conjunto diverso de enzimas classificadas como hidrolases, as quais transformam ésteres, amidas e outras funções derivadas de ácidos carboxílicos em metabólitos mais hidrofílicos. Em geral, as hidrolases são classificadas em esterases, amidases, tioesterases e fosfatases (LIMA, 2015).

Os complexos enzimáticos envolvidos no metabolismo de fase 1 evidenciam o predomínio de enzimas oxidativas, entre as quais o CYP450 tem um lugar proeminente no metabolismo hepático de fármacos das mais diversas classes terapêuticas (MONTELLANO, 2010). O CYP450 funciona como um sistema transportador de elétrons responsável pelo metabolismo oxidativo de várias substâncias endógenas (esteróides, prostaglandinas e ácidos biliares), e substratos exógenos (xenobióticos) incluindo carcinógenos, inseticidas, poluentes ambientais e fármacos (OWENS, 2006).

O metabolismo de fase 2 compreende reações de conjugação, em que os produtos formados são associados através de ligações covalentes com sustâncias endógenas, como o acido glicurônico, glutationa, sulfato ou aminoácidos, originando conjugados mais

hidrossolúveis, que são excretados preferencialmente pela urina, ou então excretados pela bile e eliminados nas fezes (MONTELLANO, 2010; LIMA, 2015).

1.1.2 Enzimas Citocromo P450

As enzimas da família citocromo P450 (CYPs) constituem uma superfamília de hemoproteínas oxidativas que catalisam a monoxigenação de um grande número de compostos endógenos e exógenos. Essas proteínas desempenham um papel chave no metabolismo de uma ampla variedade de xenobióticos, tais como fármacos, pesticidas, pre-carcinogênicos, e podem ser encontradas em uma ampla variedade de espécies. Mais de 6.500 genes de CYP450 já foram identificados, 57 deles foram encontrados no genoma humano, codificando 18 famílias e 44 subfamílias. CYP450 pode ser divido em famílias e subfamílias de acordo com a sequência homóloga de nucleotídeo, são consideradas da mesma família genética proteínas CYP que possuam > 40% de identidade na sequência de aminoácidos; enquanto que valores de identidade > 55% codificam proteínas da mesma subfamília (TAAVITSAINEN, 2001; STJERNSCHANTZ; VERMEULEN; OOSTENBRINK, 2008; SUN et al., 2012; HANDA et al., 2013; ANDRADE; SILVA; BRAGA, 2014).

A denominação do citocromo P450 tem origem nas características de hemoproteínas destas enzimas e em suas propriedades de absorção no espectro de infravermelho, observadas por Omura e Sato em 1962, em que essas enzimas apresentavam um pico de absorção máximo em 450 nm, quando reduzido na presença de monóxido de carbono (OMURA; SATO, 1964; FERNÁNDEZ et al., 2009).

Em particular, sete das cinquenta e sete isoformas humanas são responsáveis pelo metabolismo de mais de 90% de todos os fármacos prescritos no mercado. Estas são: CYP1A2, 2C9, 2C18, 2C19, 2D6, 2E1 e 3A4, sendo expressas em diferentes níveis no fígado: 1A2 (13%), 2C9 (20%), 2D6 (2%), 2E1 (7%), 2C19 (5%), e 3A4 (30%) (GUENGERICH, 2006; BERNHARDT, 2006; MONTELLANO, 2010; ZARETZKI et al., 2012).

Os membros das subfamílias exibem, em sua maioria, uma alta especificidade no metabolismo de xenobióticos, com uma ampla variedade de substratos. Algumas CYPs desempenham um papel tanto na formação e eliminação de compostos endógenos, enquanto que outras CYPs, especialmente aquelas pertencentes às famílias 1 a 3, são responsáveis principalmente pelo metabolismo de fármacos (TAAVITSAINEN, 2001).

As CYPs se encontram amplamente distribuídas em animais, plantas e protistas, e existem na natureza desde antes da divisão entre organismos eucariontes e procariontes. As

CYPs de células eucariontes possuem comprimento variável entre cerca de 480-560 aminoácidos, e são agrupadas com base na sua localização subcelular. Em procariontes, se encontram distribuídas principalmente em todas as membranas sub-celulares, sendo a mitocôndria e retículo endoplasmático as fontes mais importantes (FERNÁNDEZ et al., 2009; NELSON, 2011; LIMA, 2015).

A nível molecular o CYP450 é constituído por um grupo prostético (heme) e por uma parte protéica (Figura 1). O grupo heme consiste em um macrociclo tetrapirrólico conectado por uma ponte de metano, contendo como substituintes quatro radicais metila, dois radicais vinila e dois ácidos propiônicos. Na região central do macrociclo encontra-se o átomo de Fe bivalente ou trivalente, que está coordenado aos quatro nitrogênios do anel pirrólico e há um quinto ligante axial, que define o tipo de hemoprotéina (LIMA, 2015).

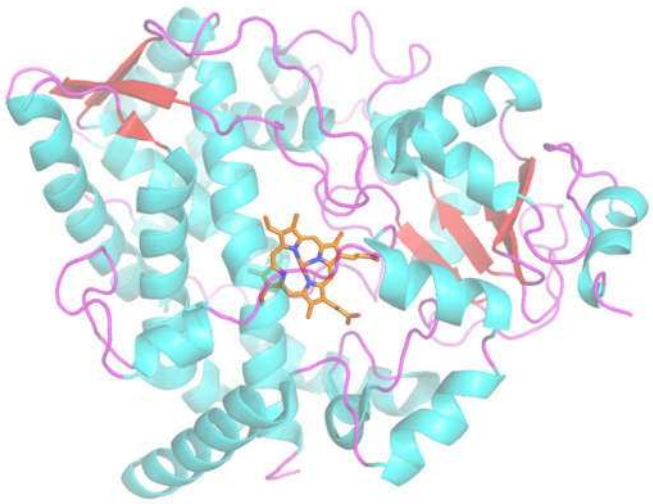

Figura 1. Representação em fita da estrutura 3D do CYP3A4. Unidade prostética representada pelo grupo heme no centro (em laranja), parte protéica pelas alfa- hélices (em azul), folhas betas (em vermelho) e alças (em rosa) (fonte, PBD:1W02).

Dentre as reações catalisadas pelo CYP450, as principais são hidroxilações, epoxidações, *N*-, *S*-, *O*-desalquilações, *N*-oxidações, sulfoxidações e desalogenações. O CYP450 catalisa a transferência de um átomo de oxigênio molecular para o substrato e o segundo átomo de oxigênio é reduzido à água utilizando o NADH (Nicotinamida adenina dinucleotídeo) ou NADPH (Nicotinamida adenina dinucleotídeo fosfato) como doador de elétrons (Equação 1)

$$\mathbf{R\text{-}H + O_2 + NADPH + H^+ \rightarrow R\text{-}O\text{-}H + H_2O + NAD\,(P^+)} \quad (1)$$

1.1.2.1 CPY3A4

A subfamília de CYP3A é expressa em altos níveis e tem uma ampla especificidade pelos substratos, é responsável por 30% do conteúdo total de P450 no fígado e é responsável pelo metabolismo de cerca de 50% dos fármacos comercializados. Portanto, a CYP3A4 é a isoenzima mais importante em termos de metabolismo de fármacos. Uma grande variedade de fármacos é capaz de se ligar a esta isoforma. Alguns exemplos de fármacos capazes de modular a atividade da CYP3A4 em diferentes categorias são: (i) Substratos, que cobrem uma ampla gama de compostos lipofílicos, tais como ciclosporina, testosterona, diazepam, midazolam (ii) indutores enzimáticos, como efavirenz, nevirapina; (iii) inibidores enzimáticos, como saquinavir, cetoconazol, cimetidina, entre outros. Vale destacar que a inibição de CYPA3A4, em muitos casos, pode conduzir a um acúmulo indesejado de outro fármaco co-administrado, aumentando o risco de causar efeitos tóxicos. A maioria das interações medicamentosas que resultaram na retirada de medicamentos do mercado pode ser rastreada através da inibição de CYP3A4 (PIRMOHAMED; PARK, 2003; ARIMOTO, 2006; MONTELLANO, 2010).

1.1.3 Inibidores enzimáticos

Atividades do CYP450 são afetadas por fatores genéticos, endógenos e ambientais, que fazem o metabolismo de fármacos ser um processo extremamente variável e individual. Esta variabilidade tem repercussões importantes para o desenvolvimento de fármacos, terapia medicamentosa clínica e em geral a sensibilidade a produtos químicos ao organismo, ou seja, xenobióticos. Dentre os fatores ambientais, compostos que causam a inibição e indução são os mais importantes (PELKONEN et al., 2008; SUN et al., 2012).

Um composto que inibe a atividade metabólica de uma enzima de CYP450 pode afetar a farmacocinética de outros substratos, resultando em uma interação fármaco-fármaco (DDIs), conduzindo potencialmente a vários efeitos colaterais e até tóxicos. Se um candidato a fármaco mostra possível capacidade de DDIs, outros estudos são requeridos pelas autoridades regulatórias para avaliar o risco em ensaios clínicos e em humanos (HANDA et al., 2013). Isto é fundamental para considerar as interações de candidatos a fármacos com as enzimas do CYP450 no processo inicial de desenvolvimento de novos fármacos (KUMAR; SURAPANENI, 2001; PELKONEN et al., 2008). Contudo, o custo e os recursos necessários para estes experimentos, limitam o número de compostos a serem testados. Neste sentido, a

capacidade em predizer a inibição de CYPs usando métodos *in silico* poderiam aumentar o número de compostos a serem testados (KUMAR; RESFSGAARD et al., 2006).

Alguns exemplos de fármacos foram proscritos do mercado devido à interação medicamentosa causada através a inibição de isoformas específicas de CYP450, tais como a terfenadina, mifebradil, cisaprida, cerivastatina, entre outros (STJERNSCHANTZ; VERMEULEN; OOSTENBRINK, 2008).

1.1.4 Indutores enzimáticos

As enzimas do CYP450 são suscetíveis à indução por xenobióticos estruturalmente diversos. O processo pelo qual a atividade destas enzimas é aumentada é denominado indução enzimática. O aumento da atividade é, aparentemente, causado por um aumento da quantidade de enzima recentemente sintetizada. A indução da enzima aumenta frequentemente à taxa de metabolismo de fármacos e diminui a duração da ação do fármaco. No entanto, a administração concomitante de dois ou mais fármacos, muitas vezes pode levar a interações medicamentosas graves como resultado da indução enzimática (JOHN M. BEALE; BLOCK, 2011).

1.2 Panorama atual da Indústria Farmacêutica

A indústria farmacêutica é responsável por pesquisar, produzir, desenvolver, comercializar e distribuir fármacos e outros produtos farmacêuticos (THOMSOM REUTERES, 2012). O mercado farmacêutico é um setor extremamente competitivo que dispõe de elevados níveis de investimento e altos riscos, e o sucesso nesse setor industrial depende de sua capacidade inovadora, mediante ao lançamento contínuo de novos produtos (DIMASI; HANSEN; GRABOWSKI, 2003; PAMMOLLI; MAGAZZINI; RICCABONI, 2011; JULIANO, 2013). Dessa forma, a indústria farmacêutica busca continuamente uma melhor posição no mercado em relação aos seus concorrentes (COHEN, 2005).

Estima-se que para o lançamento de um novo fármaco no mercado farmacêutico, são investidos em média 10 a 15 anos em Pesquisa e Desenvolvimento (P&D), com custos totais em todas as etapas de desenvolvimento de novos fármacos que podem custar até USS$ 2,6 bilhões de dólares, incluindo o custo com as falhas envolvidas no processo. Apenas 5 em cada 5.000 compostos entram na fase clínica para realizar testes em humanos, atualmente somente um recebe aprovação pelas agências regulatórias (DALKAS et al., 2013; NICOLAOU, 2014; MULLARD, 2014).

Os recentes sucessos da indústria farmacêutica são inegáveis e extraordinários. Para manter a lucratividade, a indústria farmacêutica não tem medido esforços em investir em P&D de *blockbusters*, termo em inglês utilizado para designar fármacos com vendas anuais superiores a US$ 1 bilhão (PHRMA, 2007; ADEUSI, 2011; NICOLAOU, 2014). Neste contexto, é importante mencionar o hipolipemiante atorvastatina (Liptor®, Pfizer), que só no ano de 2008 faturou extraordinários US$ 12,401 bilhões de dólares e rendeu desde seu lançamento no mercado em 1997 até 2011 aproximadamente US$130 bilhões à Pfizer (NATURE, 2011). Portanto, pode-se dizer que a capacidade da indústria em destinar recursos em P&D está diretamente relacionada à capacidade de gerar lucros (COHEN, 2005; JULIANO, 2013).

1.3 Descoberta e desenvolvimento de fármacos

A descoberta e o desenvolvimento de fármacos consiste um processo complexo e longo, que compreende várias etapas, e custos elevados. Esse processo envolve a integração de vários fatores, como inovação, conhecimento, tecnologia, gerenciamento e altos investimentos em P&D (LOMBARDINO; LOWE, 2004; PHRMA, 2013).

O processo de P&D de novos fármacos, conforme ilustrado na Figura 2 é divido em duas grandes etapas: (i) a fase pré-clínica, que envolve a pesquisa básica nos estágios iniciais, a descoberta de novas entidades químicas (NCE, do inglês *New Chemical Entities*) com potencial de aplicação clínica; e (ii) a fase clínica, em que ocorre a determinação das propriedades clínicas. Os ensaios desenvolvidos durante essas etapas visam responder a uma série de questionamentos científicos sobre um novo composto, como a avaliação de sua potência, eficácia, segurança e estabilidade (LOMBARDINO; LOWE, 2004; PHRMA, 2007).

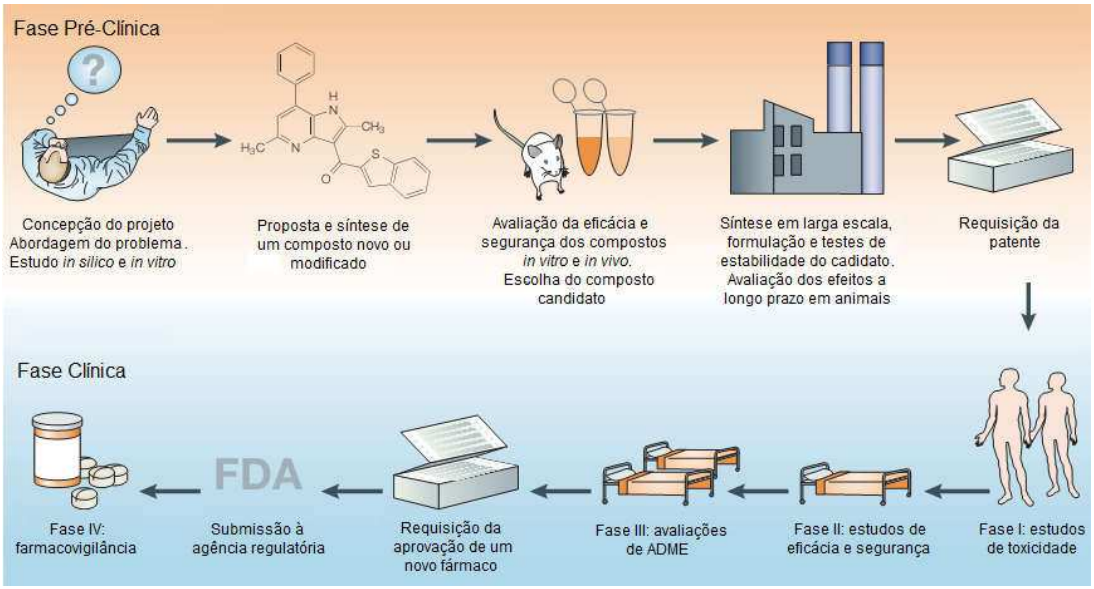

Figura 2. Etapas do processo de descoberta e desenvolvimento de novos fármacos (LOMBARDINO; LOWE, 2004).

Nos estágios iniciais da fase pré-clínica, as pesquisas se concentram na identificação e validação de alvos moleculares e na otimização de moléculas pequenas com atividade moduladora sobre o alvo biológico selecionado (Figura 3). Uma vez disponíveis e avaliadas experimentalmente quanto à capacidade moduladora do alvo proposto, as moléculas mais promissoras são selecionadas para otimização, buscando-se melhorar suas propriedades farmacocinéticas (absorção, distribuição, metabolismo e excreção, juntamente com toxicidade - ADME/Tox) após estudos *in silico, in vitro* e *in vivo* (HUGHES et al., 2011; WALTERS et al., 2011). Essa avaliação diminui tempo e custos no processo de P&D, visto que elimina candidatos com propriedades inadequadas antes de seguirem para estudos clínicos (HUGHES et al., 2011).

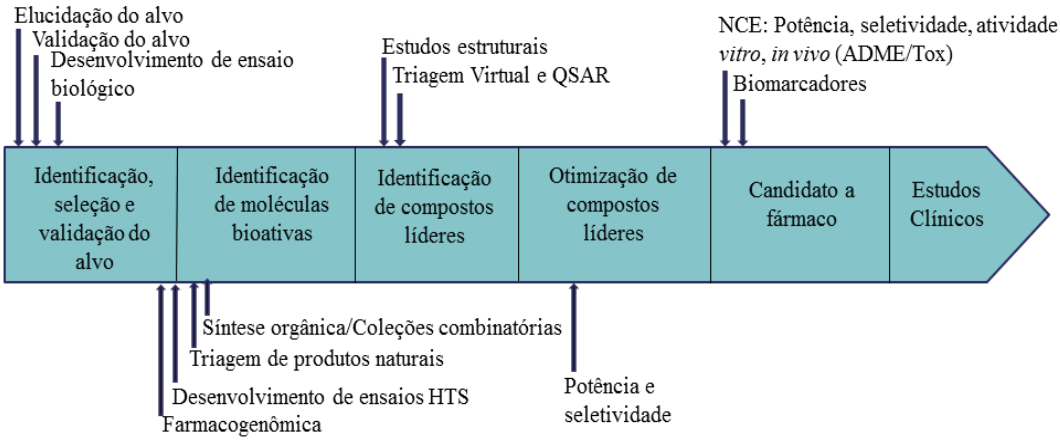

Figura 3. Identificação, seleção e otimização de novas moléculas bioativas (fase pré-clínica) (modificado de BELFIELD; DELANEY, 2006).

Após avaliação inicial do potencial terapêutico dos novos candidatos a protótipos de fármacos, as moléculas aprovadas na fase pré-clínica são submetidas às agências regulatórias, no caso dos Estados Unidos, a Agência de Administração de Alimentos e Medicamentos (*FDA*, *Food and Drug Administration*), para uma proposta de investigação de um novo fármaco (*IND*, *Invetigational New Drug*). Após a aprovação de uma nova *IND*, são iniciados os testes clínicos em humanos, para confirmação de sua segurança e eficácia. A fase clínica envolve as seguintes etapas (I-IV). Na fase I são realizados estudos de toxicidade em um grupo pequeno (20 a 100) de voluntários saudáveis. Na fase II os estudos são realizados em pacientes com a doença ou desordem (100 a 300 indivíduos) para se avaliar a dosagem, eficácia e segurança da composição. Já durante a fase III são realizadas avaliações farmacocinéticas em um grupo maior de indivíduos (300-3.000 ou mais), provenientes de regiões geográficas distintas e escolhidas randomicamente. Nesta fase o candidato a fármaco é comparado com o tratamento preconizado e determina-se sua efetividade (FDA, 2006; MCGEE, 2006). Os candidatos a fármacos aprovados ao final da fase III são submetidos à análise juntamente com a documentação dos resultados obtidos e os protocolos utilizados nos ensaios, visando obter dos órgãos reguladores a autorização para comercialização. Uma vez aprovado e inserido no mercado, têm-se início a fase IV, na qual o medicamento passa a ser comercializado e extensivamente monitorado quanto a eventuais efeitos indesejados e de longo prazo de pós- aprovação do novo fármaco (MCGEE, 2006).

O elevado risco do processo de descoberta de fármacos é evidenciado nos cerca de 80% de insucessos de NCEs que atingem as fases de desenvolvimento clínico. Das NCEs

eliminadas deste processo, 50% apresentam propriedades farmacocinéticas inapropriadas e problemas com toxicidade, conforme ilustrado na Figura 4 (VAN DE WATERBEEMD; GIFFORD, 2003).

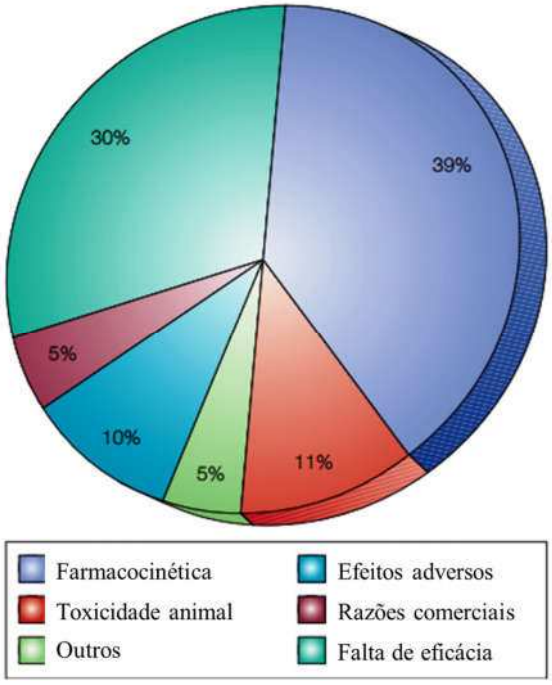

Figura 4. Principais razões de insucesso de NCEs nas fases clínicas de desenvolvimento (Adaptado de VAN DE WATERBEEMD; GIFFORD, 2003).

O conhecimento prévio das propriedades farmacocinéticas e de toxicidade (ADME/Tox) nos estágios iniciais de P&D de fármacos é capaz de fornecer informações rápidas e decisivas na otimização de novos candidatos a fármacos com elevado potencial de desenvolvimento clínico, reduzindo custos e tempo, conduzindo a um crescente interesse por novas estratégias e métodos capazes de contribuir efetivamente para o processo de otimização de propriedades de candidatos a NCEs (EKINS; HONEYCUTT; METZ, 2010).

Entre as propriedades farmacocinéticas, o metabolismo é um determinante-chave de vários processos importantes dos fármacos, tais como estabilidade metabólica, interação fármaco-fármaco, afinidade para certas enzimas metabolizadoras e toxidade de fármacos (LI et al., 2008; BRAGA; ANDRADE, 2012). O estudo do metabolismo de fármacos e imprescindível no processo de descoberta e desenvolvimento de fármacos. Apesar deste conduzir geralmente à detoxificação do organismo, algumas vezes as biotransformações metabólicas podem resultar em espécies reativas, tóxicas, em acúmulo sistêmico de metabólitos e vias de indução/inibição metabólica (KIRCHMAIR et al., 2012).

1.4 Química Medicinal e o planejamento de novos fármacos

Segundo a União Internacional de Química Pura e Aplicada (IUPAC, do inglês *International Union of Pure and Applied Chemistry*), a Química Medicinal é uma disciplina baseada na química, envolvendo aspectos das ciências biológicas, médicas e farmacêuticas, cujo objetivo é a invenção, descoberta, planejamento, identificação e preparação de compostos biologicamente ativos, o estudo do metabolismo, interpretação do mecanismo de ação a nível molecular, e das relações entre a estrutura química e atividade biológica (SAR) (BUCKLE et al., 2013).

A introdução de novas tecnologias no planejamento e desenvolvimento de novos fármacos tornou a química medicinal uma disciplina ampla em sua concepção, expandindo o seu caráter multidisciplinar, envolvendo várias áreas como farmacologia, química orgânica sintética, química computacional, entre outras (GUIDO; ANDRICOPULO; OLIVA, 2010).

Diversas estratégias da Química Medicinal podem ser empregadas no desenho molecular de novos candidatos a agentes terapêuticos. Além disso, estas estratégias são fundamentais na etapa de modificação molecular necessária à sua otimização, diminuindo efeitos colaterais, aumentando sua potência e facilitando sua interação com o receptor (BARREIRO, 2008). Na aplicação de estratégias de planejamento de fármacos, os estudos dos processos evolutivos de reconhecimento molecular em sistemas biológicos são de relevante importância, pois constituem as bases fundamentais para o entendimento de propriedades como potência afinidade, (GUIDO; ANDRICOPULO; OLIVA, 2010; GUIDO; OLIVA; ANDRICOPULO, 2011).

A integração de métodos computacionais avançados é essencial na busca de moléculas bioativas com potencial terapêutico e qualificado em uma série complexa de estudos, como a modelagem molecular, que surgiu como um conjunto de ferramentas da bioinformática que investiga as estruturas e propriedades moleculares usando a química computacional e técnicas de visualização gráfica para fornecer uma representação tridimensional próxima da estrutura real (WERMUTH; GANELLIN; LINDBERG, 1998).

Graças aos avanços obtidos em *hardware* e *software*, bem com os avanços obtidos no campo da genômica e proteômica, técnicas de cristalografia de raios-X e ressonância magnética nuclear (RMN), cerca de 87 mil estruturas 3D estão atualmente disponíveis no PDB (do inglês, *Protein Data Bank*), fornecendo, assim, a informação diferencial para aplicação das principais estratégias do planejamento de fármacos auxiliado por computador (CADD, do inglês, *Computer-Aided Drug Design*) (COHEN et al., 1996; OOMS, 2000).

O conhecimento das estruturas de alvos macromoleculares ou de complexos ligante-receptor comporta o planejamento e desenvolvimento de inibidores enzimáticos ou agonistas/antagonistas de receptores, através do processo de complementaridade molecular (estéreo/eletrostática). Neste processo, planeja-se um candidato com propriedades estruturais adequadas para o reconhecimento molecular e aumento da afinidade pelo receptor biológico. Esta estratégia é conhecida como planejamento baseado na estrutura (SBDD, do inglês *Structure-based Drug Design*) (COHEN et al., 1996).

Ao contrário, quando a estrutura 3D do alvo biológico não é conhecida, o planejamento baseia-se no estudo estrutural de uma série de ligantes conhecidos ou compostos endógenos envolvidos na fisiopatologia do processo investigado. Esta estratégia é conhecida como planejamento baseado no ligante (LBDD, do inglês *Ligand-based Drug Design*). As estratégias de SBDD e LBDD podem ser usadas conjuntamente no planejamento racional, uma vez que fornecem informações complementares (BAJORATH, 2002; WILSON; LILL, 2011; SUKUMAR; DAS, 2011).

1.5 Relações Quantitativas entre estrutura química e atividade/propriedade (QSAR/QSPR)

1.5.1 Histórico e Evolução

A metodologia de QSAR foi desenvolvida inicialmente na década de 1960, pelos trabalhos de Hansch e Fujita. Esses pesquisadores relataram que a atividade biológica de uma série de compostos pode ser linearmente correlacionada com diferentes parâmetros físico-químicos, podendo ser representado pela Equação (2).

$$Log\ (1/C) = a\pi + b\delta + cEs + d \qquad (2)$$

Em que *C* descreve a concentração molar do composto necessária para produzir uma resposta biológica definida. Os coeficientes apresentados anteriormente na Eq.2 são expressos de acordo com as propriedades estruturais, que posteriormente serão relacionados com atividade biológica na busca de um valor de log, em que π expressa o valor relativo à contribuição hidrofóbica, δ corresponde ao efeito eletrônico, *Es* corresponde ao efeito estérico, *d* é o termo independente e *a*, *b* e *c* os respectivos coeficientes determinados por análise de regressão (HANSCH; FUJITA, 1964; HANSCH, 1969).

Logo depois, Hansch propôs a extensão do modelo linear para o parabólico, quando percebeu que a resposta biológica de compostos com caráter hidrofóbico estagnava ou reduzia após atingir um valor máximo. A partir deste problema, foi percebido que compostos com alta hidrofobicidade poderiam ficar retidos nas barreiras lipofílicas, portanto têm menor probabilidade de atingir os sítios biológicos de ação (HANSCH, 1969). O modelo parabólico proposto com base numa série de dados pressupõe uma relação de segunda ordem da hidrofobicidade, expressa como o logaritmo do coeficiente de partição entre octanol e água (Log *P*), com a atividade biológica (DEBNATH, 2001), conforme descrita pela Equação 3.

$$Log\ (1/C) = a(LogP)^2 + BLogP + c \qquad (3)$$

Na década de 1970, Kubinyi desenvolveu um modelo bilinear que descreve a dependência não-linear da atividade biológica sobre o caráter hidrofóbico, no qual o termo β corresponde ao movimento do composto em sistemas multi-compartimentados, como o sistema biológico, de acordo com o descrito pela Equação 4 (KUBINYI, 1997; DEBNATH, 2001; TAVARES, 2004).

$$Log\ (1/C) = a\ Log\ P - bLog(\beta P + 1) + \ c \qquad (4)$$

Na década de 1980, os avanços em *hardware* e *software* permitiram o estudo das propriedades tridimensionais (3D) (COHEN, 1996; OOMS, 2000). Em consequência, a metodologia descrita como QSAR tridimensional (3D) surgiu. Os descritores do QSAR-3D estão relacionados ao cálculo de propriedades correspondentes à estrutura 3D dos ligantes (método independente do receptor, IR) ou do complexo ligante-receptor (método dependente do receptor, DR). Estes descritores podem ser classificados em globais (moleculares) e locais (atômicos). Os descritores globais representam uma categoria mais simples, em que os descritores (volume molecular, área molecular superficial) são derivados da estrutura molecular como um todo, enquanto que os descritores locais são derivados de fragmentos ou partes da estrutura molecular (átomos, grupos farmacofóricos) (VERLI et al., 2002).

Ainda na década de 80, Hopfinger e colaboradores introduziram a análise conformacional das estruturas e a obtenção de descritores globais 3D da forma molecular aos estudos de QSAR, criando o método denominado Análise da Forma Molecular (MAS, do inglês, *Molecular Shape Analysis*) (HOPFINGER et al., 1997). Em 1988, Cramer e colaboradores descreveram a metodologia de Análise Comparativa de Campos Moleculares

(*Comparative Molecular Field Analysis*, CoMFA), a técnica de QSAR-3D mais utilizada no mundo. Essa técnica demonstra que a propriedade biológica dos compostos pode ser correlacionada com as energias estérica e eletrostática (descritores locais) provenientes das interações formadas com o ligante no sítio ativo do alvo biológico (CRAMER; PATTERSON; BUNCE, 1988).

Em 1997, a metodologia de QSAR-4D foi introduzida por Hopfinger e colaboradores (HOPFINGER et al., 1997). Esta técnica baseia-se em na abordagem de QSAR-3D, que utiliza uma amostragem conformacional obtida através de simulação por Dinâmica Molecular (DM), reduzindo a dificuldade em identificar a conformação bioativa. Com a intenção de aperfeiçoar o poder preditivo das equações de QSAR, dimensões adicionais incorporaram-se aos métodos, como por exemplo, o Quasar QSAR-5D (VEDANI; DOBLER, 2002), no qual a teoria do encaixe induzido é considerada, ou seja, a adaptação do sítio de ligação do receptor a cada ligante individual e, mais recentemente, o Quasar QSAR-6D (VEDANI; DOBLER; LILL, 2005) que considera os modelos de solvatação simultaneamente.

Na última década, o QSAR vem apresentando inúmeros avanços e aumentando o interesse pela indústria farmacêutica e universidade, graças à inserção de vários descritores moleculares, métodos de aprendizado de máquina e parâmetros de validação (CRAMER, 2012). Com o advento das "ômicas", como genômica (CROOKE, 1998), proteômica (WILKINS et al., 1996), e metabolômica (WISHART, 2008), juntamente com o avanço de técnicas da bioinformática e de métodos de elucidação estrutural, tais como cristalografia de raios-X e ressonância magnética nuclear (RMN), tem possibilitado a elucidação estrutural e mecanismos bioquímicos de um número cada vez maior de alvos moleculares (SEOANE et al., 2013; CHERKASOV et al., 2014)

1.5.2 Aplicações

Estudos computacionais possibilitam predizer a atividade biológica e propriedades físico-químicas por meios racionais, assim como a compreensão e suposição do mecanismo de ação de séries congêneres. Sendo assim, modelos *in silico* estão sendo amplamente usados em estágios iniciais de P&D na seleção e otimização de moléculas com um grande potencial de desenvolvimento. Estes modelos contribuem para redução do custo do desenvolvimento de novos candidatos a fármacos, redução do número de animais utilizados em ensaios experimentais, promoção da química verde por aumentar a eficiência do processo de P&D,

diminuindo resíduos que seriam descartados por compostos improváveis de terem sucesso (CRONIN, 2010). Muitos esforços têm sido investidos na área científica considerado ainda emergente, o da modelagem *in silico* das propriedades ADME/Tox. As propriedades farmacocinéticas podem ser estudadas através de métodos como QSAR/QSPR. Neste âmbito, estudos de QSAR/QSPR apresentam inúmeras aplicações, tais como: (i) identificação racional de novos ligantes/protótipos com atividade/propriedade desejada; (ii) otimização da atividade/propriedade; e (iii) a identificação de compostos potencialmente perigosos em estágios preliminares do desenvolvimento (MODA et al., 2007; TROPSHA, 2010).

1.5.3 Princípios

A abordagem de QSAR/QSPR pode ser descrita como um método estatístico de análise de dados para desenvolver modelos que possam predizer corretamente determinada atividade biológica ou propriedade de compostos baseados em sua estrutura química. As técnicas de QSAR aplicam descritores baseados em estruturas moleculares e utilizam algoritmos para correlacionar os descritores obtidos com o valor da propriedade alvo de interesse (Figura 5) (MODA, 2007; GUIDO; OLIVA; ANDRICOPULO, 2011; CRAMER, 2012).

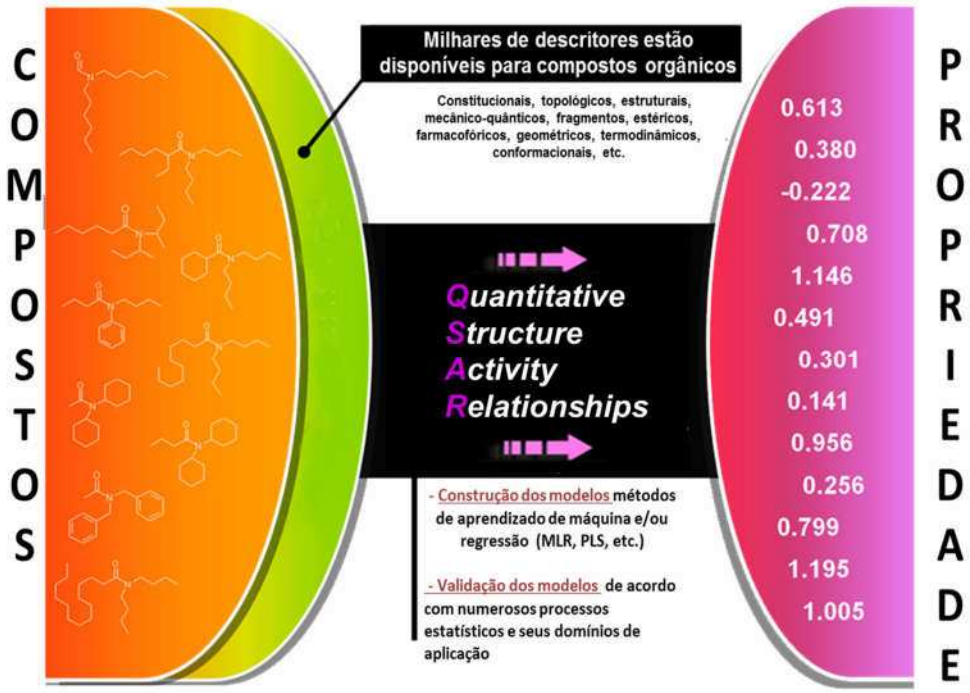

Figura 5. Processo de desenvolvimento do modelo de QSAR (adaptado TROPSHA, 2010).

Atualmente, milhares de descritores moleculares estão disponíveis e tem aumentando consideravelmente. No entanto, os descritores são muitas vezes difíceis de interpretar, sendo uma das limitações do QSAR (TODESCHINI; CONSONNI, 2008). O desenvolvimento manual de um modelo matemático considerando a quantidade de dados disponíveis atualmente seria árduo, pois seria necessário muito tempo para construir modelos com boa capacidade de predição, e com a finalidade em automatizar essa enorme quantidade de informação, métodos matemáticos e estatísticos se tornaram indispensáveis neste processo. Esses métodos, denominados de aprendizado de máquina, estabelecem peso aos descritores, ajustando a equação que relaciona a estrutura química com a atividade biológica ou propriedade (Figura 5) (TROPSHA, 2010).

1.5.4 Descritores moleculares

Nos últimos anos tem-se verificado um grande avanço no estudo das relações quantitativas entre estrutura química e atividade/propriedade (QSAR/QSPR). Primeiramente são obtidos/calculados os descritores moleculares que representam as características estruturais de uma molécula. Numa etapa seguinte, estes descritores são então correlacionados com as propriedades biológicas, a fim de se obter uma correlação quantitativa entre a estrutura e sua propriedade (RANDIC; BASAK, 2001; TODESCHINI; CONSONNI, 2008).

Um descritor molecular é o resultado final de um procedimento matemático e lógico que transforma informação química codificada em uma representação simbólica de uma molécula em um número útil ou o resultado de algum experimento padronizado. Descritores moleculares contribuem para a compreensão de propriedades moleculares e/ou podem ser utilizados na geração de um modelo matemático para a predição de determinada propriedade de outras moléculas (TODESCHINI; CONSONNI, 2008).

Os descritores moleculares são variáveis independentes que são calculados para um conjunto de dados levando informações necessárias para o desenvolvimento do modelo. Embora seja possível desenvolver modelos que usam todos os descritores, existem muitas razões para selecionar um subconjunto deles, tais como: (i) aumentar o poder preditivo dos modelos; (ii) aumentar a velocidade de trabalho para o algoritmo; e (iii) aumentar a capacidade de interpretação da relação entre descritores e atividade observada. Neste passo, o objetivo é o de reduzir a dimensão do espaço químico, sem perda de informações importante (CONSONNI; TODESCHINI, 2010).

Os descritores moleculares podem ser provenientes de dados experimentais ou calculados (teóricos ou *in silico*). Os descritores calculados podem ser classificados de acordo com a dimensionalidade da molécula usada para calculá-los, ou seja, unidimensional (1D), bidimensional (2D) e tridimensional (3D). Os descritores 1D e 2D não dependem da conformação tridimensional das moléculas, ao contrário dos descritores 3D. Este tipo de dependência ocasiona a necessidade de amostragens conformacionais e seleção da conformação bioativa que será usada na construção do modelo (TODESCHINI; CONSONNI, 2008).

Outra classificação dos descritores é com relação à natureza, que podem ser: (i) constitucionais, que são baseados nos constituintes do composto (ex., número de anéis, massa molecular, números de átomos e ligações); (i) topológicos que são obtidos a partir de grafos moleculares invariantes sem contar ligações de hidrogênio (ex., índices de Randi'c); (iii) geométricos, que são derivados de diferentes tipos de descritores baseados na geometria molecular (ex., volume molecular, área de superfície polar, entre outros); (iv) eletrostáticos, que são derivados do cálculo de cargas parciais (ex., índices de polaridade, carga parciais, entre outros); e (v) quanto-mecânicos, que são baseados nas funções de onda dos elétrons (ex., energia dos orbitais moleculares) (CONSONNI; TODESCHINI, 2010).

Para desenvolver modelos de QSAR/QSPR, os dados de atividade/propriedade são armazenados na matriz Y e os descritores na matriz X. Vários tipos de relações podem ser obtidas dessas matrizes. Modelos são então construídos a partir das variáveis independentes, com a finalidade em se obter o valor predito da variável dependente (Y), em que podem ser gerados dois tipos de variáveis aleatórias. As variáveis aleatórias que expressam resultados binários (ex., classificação) e variáveis aleatórias contínuas que são utilizados na construção de modelos de regressão (CRONIN, 2010).

Tabela 1. Representação genérica de uma matriz de dados para um estudo de QSAR/QSPR.

Identificador químico	**Atividade/ Propriedade**	**Descritor 1**	**Descritor 2**	**Descritor 3**	**...**	**Descritor n**
Molécula 1	Y_1	X_{11}	X_{12}	X_{13}	...	X_{1n}
Molécula 2	Y_2	X_{21}	X_{22}	X_{23}	...	X_{2n}
Molécula 3	Y_3	X_{31}	X_{32}	X_{33}	...	X_{3n}
Molécula 4	...	...	...	...	...	...
Molécula 5	Y_n	X_{n1}	X_{n2}	X_{n3}	...	X_{nn}

1.5.5 Métodos de aprendizado de máquina

Muitos pesquisadores enfrentaram ao longo das últimas décadas o desafio de lidar com a grande quantidade de dados químicos e biológicos de forma a identificar novos compostos biologicamente ativos (OPREA; MATTER, 2004). Nesse sentido, a área da química que utiliza técnicas computacionais para modificar dados em informação e informação em conhecimento, com a intenção de aprimorar a identificação e otimização de compostos no processo de desenvolvimento de fármacos é denominado de quimioinformática (XU; HAGLER, 2002; VARNEK; BASKIN, 2011). Com a finalidade em lidar com esse problema, métodos de aprendizado de máquina têm sido utilizados e são tão importantes quanto os dados de atividade biológica e descritores moleculares (MELVILLE; BURKE; HIRST, 2009).

Aprendizado de máquina (AM) é um ramo da inteligência artificial (IA) que introduz conceitos e métodos para a construção de modelos que irão possibilitar a extração de informações a partir de um determinado conjunto de dados (MITCHELL, 1997). Esse campo de pesquisa possibilita o reconhecimento de padrões que estão dispostos em forma de informações nos dados extraídos de um determinado conjunto populacional ou amostral. Esse aprendizado é construído em três estágios: (i) representação dos dados; (ii) otimização da hipótese; (iii) generalização (TROPSHA, 2011). Métodos de aprendizado de máquina são uma poderosa ferramenta para resolução de problemas que envolvem o metabolismo de substâncias químicas. Grande parte dos modelos de QSAR para o metabolismo de fármacos está voltada para a predição da interação com CYP450 para novas entidades químicas (NCEs) (ARIMOTO, 2006; MISHRA; TRIPATHI; YADAV, 2010; WELLING, 2010).

Os métodos de aprendizado de máquina podem ser divididos de acordo com o problema de aprendizagem que serão impostos para a construção de um modelo. A resolução para esta problemática é o delineamento de acordo com as informações que são contidas em um conjunto de dados, ou seja, é realizada a divisão do conjunto de dados em dados rotulados e dados não rotulados. Esta divisão é feita com a finalidade de otimizar a escolha de um algoritmo de aprendizado de maquina (MITCHELL, 1997; KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

Os dados rotulados são descritos como um conjunto de dados que dispõem tanto informações de entrada como de saída. Nos modelos de QSAR estes dados são representados pelas estruturas químicas e atividade biológica respectivamente, fornecendo assim uma classe rotulada. Contudo, na literatura existem conjuntos de dados que não contemplam ambas as

informações de entrada e saídas, estes dados são denominados dados não rotulados (WITTEN; FRANK; HALL, 2011).

Nos dados rotulados, as classes são previamente estabelecidas pelas informações de entrada e saída o que garante a utilização de métodos supervisionados para modelagem destes dados. Sendo assim, um algoritmo de AM supervisionado utiliza dados rotulados para gerar um classificador que irá contribuir para a alocação de dados não utilizados na construção do modelo, ou seja, para a classificação de novos dados não utilizados na construção do modelo.

Em alguns conjuntos de dados não há presença de informações que forneçam dados rotulados na forma de entrada e saída, como mencionado anteriormente. Isto dificulta o enriquecimento de informações necessárias para serem utilizados por algoritmos de aprendizado de máquina supervisionado. Portanto, a modelagem de dados não rotulados a partir de algoritmos supervisionados, favorece a ocorrência de modelos com baixo poder preditivo ou *underfitting*. Para se alcançar ótimos modelos utilizando esses dados, é necessário a utilização de métodos não supervisionados, que promovam uma indução das classes de acordo com os dados não-rotulados (BEN-DOR et al., 2000).

Além disso, é de suma importância um conjunto de dados rotulados para construção de modelos que relacionam a estrutura química e a propriedades biológicas. Possibilitando a geração de uma função de classificação capaz de predizer novos dados de saída a partir de novos dados de entrada (estrutura química).

Portanto, a premissa em se utilizar métodos de AM em estudo de QSAR/QSPR, está centrada na compilação de um algoritmo capaz de ser "treinado" para gerar modelos que sejam capazes de discriminar compostos ativos/inativos (modelos binários) e a predições numéricas da propriedade biológica (modelos contínuos).

1.5.6 Boas práticas de desenvolvimento e validação em QSAR

A validação é um aspecto crucial de qualquer estudo de QSAR. É o processo pelo qual a confiabilidade e relevância de um processo são estabelecidas para um propósito específico (VEERASAMY et al., 2011). Várias diretrizes e recomendações de boas práticas de desenvolvimento e validação de modelos de QSAR foram publicadas na última década (CHERKASOV et al., 2014; TROPSHA, 2010). A OECD (*Organization for Economic Co-operation Development*) publicou no ano de 2004 alguns princípios que foram estabelecidos para validação de modelos de QSAR para uso prático de agências de regulamentação. No futuro, a aceitação do QSAR como uma fonte alternativa para dados não testados, será

baseado na confiabilidade e transparência de um determinado modelo de QSAR dentro de um enquadramento regulamentar específico. O QSAR além de ser um método que pode reduzir o número de reagentes e compostos a serem testados nas pesquisas, também permite priorizar o número de compostos a serem testados e diminuir os gastos com reagentes. O uso do QSAR pode ser aplicado em estudos ambientais, estudos de toxicologia ocupacional, entre outros, o que pode afetar a economia do País. Os princípios são: (i) definir endpoint (atividade biológica ou propriedade); (ii) algoritmo claro; (iii) domínio de aplicabilidade (DA) definido; (iv) avaliação apropriada da robustez e preditividade; (v) e interpretação mecanística, que significa encontrar relações entre os descritores e a atividade biológica ou propriedade, com o intuito de se compreender melhor o mecanismo de ação de uma estrutura química ou aprofundar o conhecimento biológico sobre a propriedade em estudo (OECD, 2004).

1.5.6.1 Preparo do conjunto de dados

Com relação aos princípios estabelecidos pela OECD, outros elementos também são considerados fundamentais durante o desenvolvimento e validação de modelos de QSAR/QSPR e estão representados na Figura 6.

Figura 6. Fluxograma para o preparo de conjunto de dados químicos (modificado de FOURCHES, MURATOV E TROPSHA, 2010).

O conjunto de dados utilizado na construção dos modelos de QSAR/QSPR deve ser preparado cuidadosamente, com a premissa de evitar qualquer tipo de erro que possa vir a interferir na qualidade do modelo (TROPSHA, 2010). Durante esse processo, compostos que não podem ser tratados adequadamente por meio das técnicas de quimioinformática usadas no estudo, precisam ser removidos do conjunto de dados. Uma vez que esses compostos permaneçam no conjunto de dados, eles podem conferir informações negativas durante o desenvolvimento dos modelos.

Alguns quimiotipos específicos, tais como anéis aromáticos e grupos nitro, e formas tautoméricas precisam ser padronizados e os contra-íons removidos, para impedir que a mesma estrutura permaneça no conjunto de dados. Além disso, duplicatas necessitam ser identificadas e removidas. Uma inspeção manual (visual) é exigida ao final do processo para garantir que todas as estruturas conferidas estejam corretas (FOURCHES; MURATOV; TROPSHA, 2010).

1.5.6.2 Validação dos modelos

A confiabilidade de um modelo de QSAR/QSPR depende de sua capacidade de generalização, ou seja, de predizer corretamente, com alta taxa de acerto, determinada propriedade biológica de compostos que não foram utilizados durante o processo de modelagem. Essa capacidade de generalização é denominada preditividade. Existem dois métodos principais que são utilizados para determinar a capacidade preditiva dos modelos de QSAR, são eles: validação interna e validação externa. Para a validação interna o conjunto treinamento que é utilizado para construção dos modelos de QSAR. Entretanto, na validação externa é utilizado um conjunto teste, o qual não foi usado durante a construção dos modelos, juntamente com o emprego de vários parâmetros estatísticos (TROPSHA; GOLBRAIKH, 2007; GRAMATICA, 2007).

1.5.6.3 Definição do domínio de aplicabilidade (DA)

Teoricamente, modelo de QSAR pode predizer a propriedade biológica de qualquer composto químico. Contudo, se sua estrutura química for muito diferente das estruturas usadas na geração do modelo, essa predição pode ser incorreta, visto que o modelo não cobre o espaço químico no qual se insere o composto em questão. É recomendável que o DA de um modelo de QSAR seja descrito em termos dos parâmetros mais relevantes, ou seja, geralmente

os descritores utilizados no modelo. O conhecimento do DA assegura a qualificação da confiabilidade do modelo na predição de novos compostos. Vale ressaltar que a determinação do DA é obrigatória para aceitabilidade de um modelo de QSAR (TROPSHA, 2010).

1.5.7 Revisão bibliográfica de modelos de QSAR para predição do metabolismo de fármacos

Apesar de muitos modelos de QSAR para a predição do metabolismo já estarem descritos na literatura (Tabela 2), e aparentemente estarem bem ajustados e serem robustos, uma análise crítica dos mesmos revela vários problemas importantes.

Os dados referentes aos parâmetros estatísticos são de suma importância na motivação para geração de modelos com maior capacidade de predizer a propriedade biológica frente à metabolização de xenobióticos pelas enzimas do citrocromo P450. Além disso, o uso de parâmetros como domínio de aplicabilidade e randomização da variável Y, confere maior robustez e capacidade preditiva, garantindo a confiabilidade dos modelos gerados. A randomização da variável Y é uma ferramenta usada na validação de modelos de QSAR, em que o desempenho do modelo original nos dados descritos é comparado com as respostas dos modelos construídos pela randomização da propriedade biológica (selecionadas aleatoriamente). Caso a preditividade dos modelos com variável Y aleatória obtenha resultados melhores que os modelos de QSAR, então os modelos devem ser descartados, visto que os valores devem estar próximo ao aleatório, que compreende a 0,50 (RÜCKER; RÜCKER; MERINGER, 2007).

Vários modelos de QSAR reportados na Tabela 2 não obedeceram aos critérios conforme as orientações OECD (2004) para o desenvolvimento e validação de modelos de QSAR. Como disponível na Tabela 2, a maioria dos estudos publicados não tem definição do DA e nem robustez comprovada do teste de randomização da variável Y.

Apesar de serem reportados inúmeros estudos na literatura apresentando bons valores estatísticos, muitos destes estudos não proporcionam uma predição confiável para o metabolismo de fármacos. A falta de confiabilidade dos modelos gerados nestes estudos pode ser atribuída pela falta de métricas como randomização da variável Y e o DA. O uso destas métricas nos estudos de QSAR assegura a confiabilidade da predição dos modelos gerados. Além de ser um critério estabelecido pela OECD, a utilização destas ferramentas na construção de modelos de QSAR.

Tabela 2. Estudos reportados de QSAR para metabolismo mediado por CYP450.

Tipo de ensaio biológico	Número de compostos utilizados	Fonte do conjunto de dados	Descritores utilizados	Método estatístico para geração de modelos	Resultados dos modelos	DA	Randomização de Y	Referência
Inibição enzimática Ensaio/ K_m	52	Vários	(MS)-WHIM	GFA	r^2=0.69 q^2=0.58	Não	Não	(SNYDER et al., 2002)
Inibição enzimática Ensaio/ K_m	39	Vários	CoMFA	PLS	r^2=0.453-0.743 *s* = -0.002-0.371 *SEE*=0.21-0.005	Não	Não	(HAJI-MOMENIAN et al., 2003)
Inibição enzimática Ensaio/ IC_{50}(μM)	36	Vários	CoMSIA	PLS	q^2=0.60-0.70 r^2=0.70-0.90 *SEE*=0.40-0.65 *SEP*=0.75-0.80	Não	Não	(VAZ et al., 2005)
Inibição enzimática Ensaio/ IC_{50}(μM)	42	(RAHNASTO et al., 2005)	CoMFA	PLS	q^2=0.50-0.52 r^2=0.83-0.87 S_{PRESS}=0.50-0.69	Não	Não	(RAHNASTO et al., 2005)
Inibição enzimática Ensaio/ K_i Substratos	1.404	Vários	CONCORD, DRAGON	C-SVM, GA	*Conc*.=92.0-97.0 *MCC*=0.742-0.899 *SE*=75.0-98.2 *SP*=90.0-100.0	Não	Não	(YAP; CHEN, 2005)

Tabela 2. Continuação**.**

Tipo de ensaio biológico	Número de compostos utilizados	Fonte do conjunto de dados	Descritores utilizados	Método estatístico para geração de modelos	Resultados dos modelos	DA	Randomização de Y	Referência
Ensaio de ligação de microssomo utilizando fração microssomal de fígado de rato	18	Vários	Físico-Químicos	MLR	F=115-191 r=0.980-0.988 s=0.209-0.169	Não	Não	(ITOKAWA et al., 2006)
Inibição enzimática Ensaio/ IC_{50}(μM)	40	Vários	CoMFA	PLS	q^2=0.71 r^2=0.85 r^2_{pred}=0.80 S_{PRESS}=0.64	Não	Não	(KORHONEN et al., 2007)
Michaelis constants (Km)	379	Vários	Topológico	ANN, DT, k-NN, RBF, SVM	SE=76.2-95.5 SP=64.0-94.1	Não	Não	(TERFLOTH; BIENFAIT; GASTEIGER, 2007)
Ensaio de inibição enzimática/ IC_{50}(mM)	33	Vários	MOE	MLR	r^2=0.522-0.907	Não	Não	(APPIAH-OPONG et al., 2008)
Inibição enzimática Ensaio/ IC_{50}(μM)	42	(RAHNASTO et al., 2005)	Eletrônico, espacial, Shape e termodinâmico	GFA, G/PLS	r^2=0.561– 0.898 r^2_a= 0.508-0.870 R^2_{pred}=0.615-0.914,r^2_m=0.494–0.876	Não	Não	(ROY; ROY, 2008)

Tabela 2. Continuação.

Tipo de ensaio biológico	Número de compostos utilizados	Fonte do conjunto de dados	Descritores utilizados	Método estatístico para geração de modelos	Resultados dos modelos	DA	Randomização de Y	Referência
Ensaio de inibição enzimática/ IC_{50}	28	(LEWIS; LAKE; DICKINS, 2006)	Eletrônico, espacial, topológico e termodinâmico	ANN, MLR, FA-MLR, GFA, G/PLS,	q^2=0.644-0.836 r^2=0.590-0.916 r^2_m=0.496-0.735 $r^2_{m(LOO)}$=0.4850.771 r^2_{pred}=0.573-0.701	Não	Sim	(ROY; PRATIM ROY, 2009)
Inibição enzimática Ensaio/ K_i	20	Vários	CoMFA	PLS	*EL*=39.6 *F*=84.707 q^2=0.728 r^2=0.941 *Spress*=0.679 *s*=0.317 *ST*=60.4	Não	Não	(YASUO et al., 2009)
Ensaio de inibição/ IC_{50}	7.679	Vários	Fingerprints, 2D	NN, PLS	r^2=0.38-0.89 *S.E*=0.35-0.89 *S.D*=0.71-1.03	Não	Não	(EWING; FEHER, 2010)
Afinidade de ligação	226	DrugBank 2.5	*ADMEWORKS, CDK* e *TSAR.*	BayesNet, IB1, RF, RoF SVM	*AC*=70.80-86.60 *MCC*=0.15-0.63 *SE*=52.63-81.08 *SP*=72.46-85.92	Não	Não	(MISHRA; AGARWAL; RAGHAVA, 2010)
Afinidade de ligação	670	Vários	Físico-Químicos, estérico e topológico	NaiveBayes, DT, SVM	*AC*=53.7-85.6 *SE*=40.0-94.8	Não	Não	(CARBON-MANGELS; HUTTER, 2011)

Tabela 2. continuação.

Tipo de ensaio biológico	Número de compostos utilizados	Fonte do conjunto de dados	Descritores utilizados	Método estatístico para geração de modelos	Resultados dos modelos	DA	Randomização de Y	Referência
Ensaio de inibição de microssomo utilizando fração microssomal de fígado humano	51	Vários	Constitucional, eletrostático e topológico.	MLR	q^2=0.30-0.57 r^2=0.55-0.75 r_o^2=0.64-0.65	Não	Não	(SARACENO et al., 2011)
Ensaio de inibição de microssomo utilizando fração microssomal de fígado de humano/ Inibição enzimática	2.865	Yap, C. W.; Chen, Y. Z. J. Chem. Inf. Model. 2005, 45, 982.	Físico-Químicos	PLS	*Con*=68.0-86.3 *SE*=50.0-78.0 *SP*=73.6-90.9	Sim	Não	(JÓNSDÓTTIR et al., 2012)
Afinidade de ligação	2.870	Vários	Fingerprints, Topológico.	*k*-NN, SVM	*AC*=85.0-75.0 *AUC-ROC*=0.7-0.84 *Con*=64.8-77.1 *SE*=72.0-82.8 *SP*=64.0-76.7	Sim	Não	(SUN et al., 2012)
Ensaio enzimático/ K_i	309	Vários	Fingerprints	GG, GE, JJ	*AC*=85.3-47.1	Não	Não	(BURTON et al., 2013)
Ensaio de inibição enzimática/ K_i e IC_{50}	34	Vários	CoMFA	PLS	q^2=0.565-0.699 r^2=0.568-0.986 *s*=0.185-0.197 S_{press}=0.696-0.804	Não	Não	(HANDA et al., 2013)

Tabela 2. continuação.

Tipo de ensaio biológico	Número de compostos utilizados	Fonte do conjunto de dados	Descritores utilizados	Método estatístico para geração de modelos	Resultados dos modelos	DA	Randomização de Y	Referência
Afinidade de ligação	3.249	Vários	VolSurf	C-SVM, *One-classe*-SVM	*Pre*=99.1-9.0 *SE*=25.0-99..3 *SP*=66.2-98*G-mean*=0.17-0.94	Não	Sim	(MARTINEZ-SANZ et al., 2013)
Ensaios de qHTS para inibição CYP 3A4 utilizando enzima luciferase	70	PubChem: 844	1D, 2D	LinBiExp	r^2=0.51-0.97 *s*=0.71-0.89	Não	Não	(BUCHWALD; YAMASHITA, 2014)
Inibição enzimática/ K_i e IC_{50}	16	Vários	CoMFA, CoMSIA	PLS	*F*=0.687-0.795 q^2=0.687-0.795 r^2=0.948-0.962	Não	Não	(SHITYAKOV et al., 2014)

AC: *Acurácia*; q^2: *valores de validação cruzada;* r^2: *coeficiente de determinação*; r^2 pred: *coeficiente de determinação predita - conjunto de validação*; r:*coeficiente de correlação*; r^2_a: *variância explicada*; r^2_m: *Média dos valores de coeficiente de corelação*; Spress: *Desvio padrão do erro predito*; RSMEP: *Média da Raiz Quadrada do erro predito*; F; F value; EL: *Valores de contribuição estroestática*; ST: *Valores de contribuição estérica*; s: *Desvio padrão*; SD: *Erro padrão*; MCC: *Coeficiente de Correlação de Matthews*; DA: *Domínio de Aplicabilidade*; ES: Especificidade; SE: *Sensibilidade; qHTS: Quantitaive High-Throughput Screnning;* CIC: *Inibidores preditos corretamente;* CNIC: *Inibidores preditos incorretamente*; CoMFA: *A comparative molecular Field analysis*; CoMSIA: GFA*: Genetic Function Approximation; A comparative molecular similarity index analysis*; PLS: *Partial leasts quares*; GG: *Algorithm Johnson*; GE: *Johnson Algorithm Exhaustive*; GA*: Genetic Algorithm*; DT: *Decision Tree;* RBF: *The Radial Basis Function*; FA*: Factor Analysis;* LDA: *linear discriminant analysis*; LR: *logistic regression*; KNN: *k-Nearest Neighbors*; GKW: *Gaussian Kernel Weighted*; MLR: *multiple linear Regression*; RF*: Randon Forest*; ANN: *artificial neural networks*;; C-SVM: *circular support vector* machine; SVM: *support vector machine*; Conc: Concordance; I_{OH}: *Compostos com e sem grupo hidroxila;* E_{HOMO}: *Nível de energia do Homo*; TFN: *Taxa de Falsos Negativos*; TFP: *Taxa de Falsos Positivos*; BayesNet: *Bayesian Network*; IB1: *Nearest neighbor classifier*; RoF: LinBiExp: *linearized biexponential model*; SEE: *Erro padrão estimado.*

As principais falhas dos modelos apresentados na Tabela 2 são (i): falta de determinação do DA: (SNYDER et al., 2002; HAJI-MOMENIAN et al., 2003; YAP; CHEN, 2005; RAHNASTO et al., 2005; VAZ et al., 2005; ITOKAWA et al., 2006; JENSEN et al., 2007; KORHONEN et al., 2007; TERFLOTH; BIENFAIT; GASTEIGER, 2007; ITOKAWA; YAMAUCHI; CHUMAN, 2007; APPIAH-OPONG et al., 2008; YASUO et al., 2009; ROY; PRATIM ROY, 2009; MISHRA; AGARWAL; RAGHAVA, 2010; EWING; FEHER, 2010; DIDZIAPETRIS et al., 2010; SARACENO et al., 2011; CARBON-MANGELS; HUTTER, 2011; HANDA et al., 2013; BURTON et al., 2013; MARTINEZ-SANZ et al., 2013; BUCHWALD; YAMASHITA, 2014; SHITYAKOV et al., 2014); (ii) não foi comprovado o grau de randomização pelos valores do teste de randomização da variável Y (SNYDER et al., 2002; HAJI-MOMENIAN et al., 2003; YAP; CHEN, 2005; RAHNASTO et al., 2005; VAZ et al., 2005; ITOKAWA et al., 2006; JENSEN et al., 2007; KORHONEN et al., 2007; TERFLOTH; BIENFAIT; GASTEIGER, 2007; ITOKAWA; YAMAUCHI; CHUMAN, 2007; APPIAH-OPONG et al., 2008; YASUO et al., 2009; ROY; PRATIM ROY, 2009; MISHRA; AGARWAL; RAGHAVA, 2010; EWING; FEHER, 2010; DIDZIAPETRIS et al., 2010; SARACENO et al., 2011; CARBON-MANGELS; HUTTER, 2011; JÓNSDÓTTIR et al., 2012; SUN et al., 2012; HANDA et al., 2013; BURTON et al., 2013; MARTINEZ-SANZ et al., 2013; BUCHWALD; YAMASHITA, 2014; SHITYAKOV et al., 2014). Além disso, a maioria dos modelos relatados na literatura utilizou conjuntos de dados pequenos para a geração dos modelos de QSAR, o que pode ser um problema no que diz respeito à extrapolação para predição de compostos que estão fora do domínio de aplicabilidade.

Outro ponto negativo a ser ressaltado com os dados descritos na Tabela 2 é a ausência da diversidade estrutural presente nos conjunto de dados. Appiah e colaboradores realizaram um estudo de QSAR utilizando um conjunto de dados de 33 análogos de cumarina, com finalidade em predizer a atividade inibitória para as isoformas de CYP1A2, CYP3A4, CYP2B6, CYP2C9 e CYP2D6 (APPIAH-OPONG et al., 2008). Em outro trabalho, Itokawa e colaboradores realizaram um estudo com 18 compostos azóis para avaliar a afinidade de ligação com as isoformas de CYP2B e CYP3A4 (ITOKAWA; YAMAUCHI; CHUMAN, 2007). Outro estudo publicado por Itokawa e colaboradores conduziu a compilação de modelos de inibição de CYP2D6 e 3A4 para 18 compostos azóis com atividade fungicida (ITOKAWA et al., 2006). Estes estudos são restritos para algumas classes químicas na predição do metabolismo de fármacos, pois apresentam restrição química. Contudo, é necessário a construção de modelos que possam classificar um número maior de compostos

estruturalmente diversos. Isto irá contribuir para o aumento do espaço químico dos compostos, o que consequentemente permite uma maior confiabilidade da predição de um conjunto de dados estruturalmente diverso.

2 JUSTIFICATIVA E OBJETIVOS

2.1 Justificativa e Objetivo Geral

O estudo do metabolismo de fármacos constitui uma etapa crucial durante o seu desenvolvimento, sendo necessário para a aprovação para uso em humanos. Diversos métodos computacionais têm sido desenvolvidos para acelerar o processo de desenvolvimento de novos fármacos, inclusive para predizer o metabolismo nos estágios iniciais deste processo. No entanto, os estudos de QSAR para predição do metabolismo reportados na literatura não obtiveram sucesso em prover à comunidade modelos robustos capazes de avaliar o metabolismo de compostos químicos candidatos a fármacos.

Diante do exposto, este trabalho teve como objetivo central o desenvolvimento de modelos de QSAR robustos e preditivos utilizando o maior conjunto de dados disponível na literatura, visando identificar substratos e inibidores de CYP3A4, responsável pelo metabolismo de novos compostos candidatos a fármacos e disponibilizar os modelos para uso da comunidade científica.

2.2 Objetivos Específicos

- Integrar, preparar e balancear conjuntos de dados de substratos e inibidores de CYP3A4 reportados na literatura;
- Gerar e validar modelos de QSAR de substratos e inibidores de CYP3A4;
- Interpretar os modelos desenvolvidos para compreender a relação estrutura atividade (SAR) pelo qual os fármacos são metabolizados e/ou inibem a CYP3A4;
- Desenvolver um servidor para a predição *in silico* de perfis de CYP responsável pelo metabolismo de novos compostos e disponibilização na web.

3 MATERIAIS E MÉTODOS

3.1 Conjunto de dados

3.1.1 Conjunto de dados de substratos de CYP3A4

O conjunto de dados de substratos de CYP3A4 utilizado neste trabalho consiste em 8.214 compostos, dos quais 475 são substratos de CYP3A4 (ZARETZKI et al., 2012). Os compostos restantes são não-substratos (inativos) para CYP3A4 (total de 7.739), que foram retirados do banco de dados *PubChem bioassay* (ID:1851).

3.1.2 Conjunto de dados de inibidores de CYP3A4

O conjunto de dados de inibidores de CYP3A4 é o maior conjunto de dados de domínio público para CY3A4, contendo 42.295 compostos apresentando vários tipos de propriedade biológica, como inibidores, substratos e não substratos. O conjunto de dados foi obtido do ensaio ChEMBL340.

3.2 Preparo do Conjunto de dados

Informações das estruturas de compostos foram retiradas das bases de dados PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChEMBL (https://www.ebi.ac.uk/chembl/) e da literatura (ZARETZKI et al., 2012) usando-se o número de registro do Chemical Abstracts Services (CAS) ou nomes químicos. Compostos que não possuem estrutura química definida/depositada em uma dessas bases de dados foram removidos. Cada conjunto de dados foi cuidadosamente preparado de acordo com a metodologia proposta por (FOURCHES; MURATOV; TROPSHA, 2010)

Com o intuito de automatizar o pré-processamento de dados, foi utilizado um protocolo desenvolvido em nosso laboratório (*KSAR)* implementado na plataforma KNIME, como apresentado na Figura 7. Quimiotipos específicos como anéis aromáticos e grupos nitro (RNO_2), e formas tautoméricas foram padronizados e os contraíons removidos. A presença de duplicatas, ou seja, compostos idênticos reportados mais de uma vez no mesmo conjunto de dados é conhecida por conduzir a um sobre-ajuste dos modelos de QSAR. No entanto, a análise de tais registros também permite estimar a qualidade do conjunto de dados, isto é, se os dados de atividade para o mesmo composto reportado mais de uma vez são semelhantes, a qualidade dos dados é alta; porém, se existe um grande desvio dos valores experimentais, a

qualidade dos dados é baixa. As duplicatas foram identificadas usando o software HiT QSAR (KUZ'MIN; ARTEMENKO; MURATOV, 2008) e ISIDA Duplicates, e cuidadosamente analisadas. Quando os valores experimentais associados a um mesmo composto eram idênticos, apenas um dos registros era mantido no conjunto de dados. Caso contrário, todos os registros eram removidos.

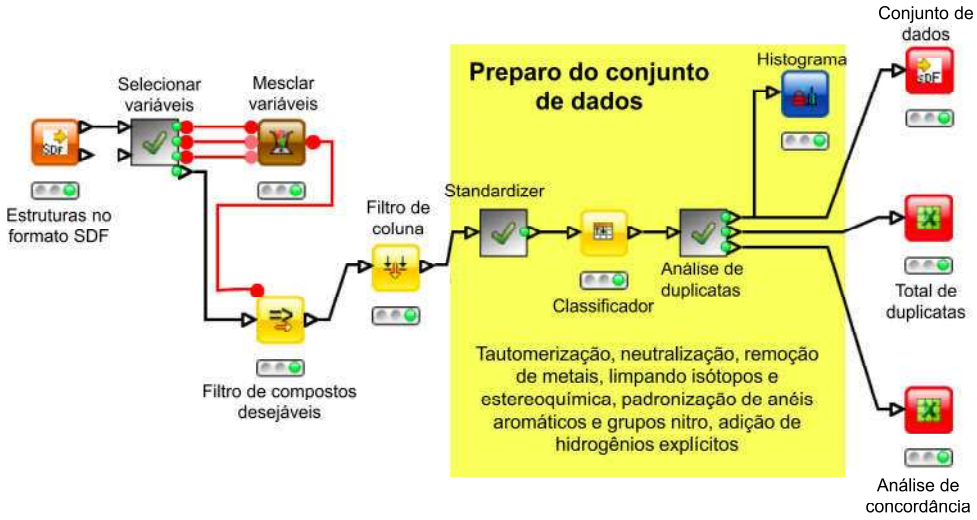

Figura 7. Protocolo *in house* (KSAR) para o preparo e padronização do conjunto de dados

3.3 Cálculos dos descritores moleculares

Quatro tipos diferentes de descritores moleculares *fingerprints* foram calculados e utilizados como variáveis independentes neste estudo, refletindo a ausência (0) ou presença (1) de fragmento substrutural de cada composto (DUAN et al., 2010).

3.3.1 *MACCS* (*The Molecular ACCess System*)

Os descritores MACCS foram calculados utilizando o RDKit (http://www.rdkit.org) na plataforma KNIME. Os descritores MACCS são compostos por uma coleção de 166 subestruturas predefinidas associadas a um padrão SMARTS. Esses descritores foram planejados para pesquisas de subestrutura; assim, eles são muitas vezes utilizados como descritores para estudos de avaliação comparativa (TODESCHINI; CONSONNI, 2009).

3.3.2 *FeatMorgan*

Os descritores FeatMorgan são fragmentos circulares baseadas no algoritmo Morgan (*FCFP-like*). Eles combinam o algoritmo *fingerprint* Morgan RDKit com características

farmacofóricas. O farmacóforo é o conjunto de características estéricas e eletrônicas essenciais para a interação com o alvo biológico e FCFPs são impressões digitais topológicas circulares onde cada bit inicial representa um farmacóforo. Um número de interações é realizado para combinar os farmacóforos iniciais identificados com os farmacóforos dos vizinhos até atingir um diâmetro especificado. A regra FCFP é derivada a partir de definições farmacofóricas (por exemplo, doador, aceptor, aromático, halogênio, base, ácido, etc.) (MORGAN, 1965; YANG, 2010).

3.3.3 *PubChem*

Descritores PubChem foram calculados utilizando o CDK também implementado na plataforma KNIME. PubChem *fingerprints* consistem em um vetor com 881 bits que representam a ausência (0) ou a presença (1) de uma subestrutura (fragmento) para cada composto. A representação de estruturas químicas 2D é baseada em elementos específicos, como por exemplo: tipos de anel, ligações simples ou duplas, ambiente atômico (vizinhos mais próximos), entre outros (STEINBECK et al., 2003).

3.3.4 *Atom Pair* (AP)

O descritor AP codifica todos os pares de átomos na molécula junto com o comprimento de ligações mais curtas e pelo caminho de ligação entre eles. Para cada par de átomos é atribuído uma das cinco classes: doador de ligação de hidrogênio, aceptor de ligação de hidrogênio, carregado positivamente, carregado negativamente e lipofilicidade (CARHART; SMITH; VENKATARAGHAVAN, 1985; BASKIN; VARNEK, 2008).

3.4 Geração e otimização dos modelos de QSAR

Os modelos de QSAR foram gerados utilizando um fluxo de trabalho desenvolvido em nosso laboratório, implementado na plataforma KNIME, em que vários protocolos foram executados. O fluxo de trabalho para geração dos modelos de QSAR incluiu três grandes passos (CHERKASOV et al., 2014; TROPSHA, 2010): *(i)* preparo/análise do conjunto de dados (cálculo e seleção de descritores e compostos), *(ii)* construção dos modelos e *(iii)* validação/seleção dos modelos.

Na Figura 8 está representado um fluxograma geral da construção dos modelos de QSAR de substratos e inibidores de CYP3A4 utilizado neste trabalho.

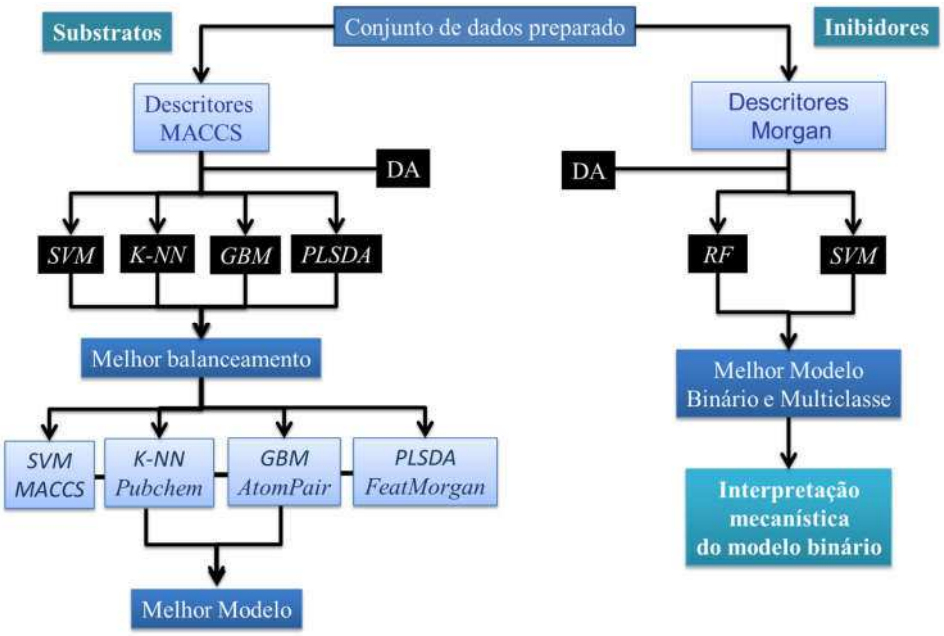

Figura 8. Fluxograma geral da construção dos modelos de QSAR para substratos e inibidores de CYP3A4 utilizado neste trabalho.

Empregou-se o método de validação cruzada externa de 5-*fold* (Figura 9), em que o conjunto de dados foi dividido pelo algoritmo *Kennard-Stone* em cinco subgrupos de tamanhos iguais. Esse algoritmo faz uma divisão racional do conjunto de dados, calculando a distância euclidiana dos compostos do conjunto modelagem e de validação externa, garantindo que os compostos do conjunto modelagem sejam distribuídos de forma uniforme em toda área ocupada pelos compostos representativos das proximidades do conjunto de validação externa.

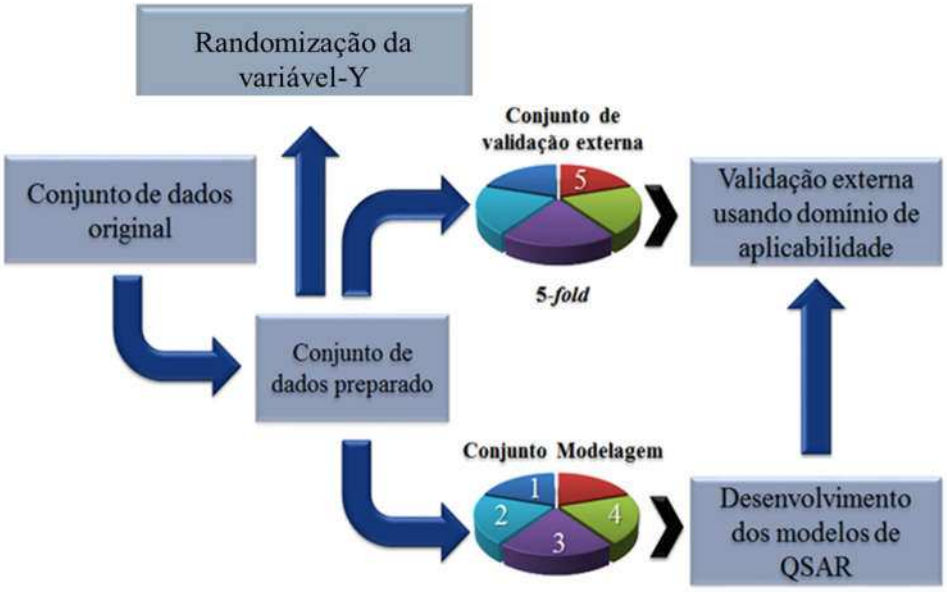

Figura 9. Esquema do fluxo de trabalho com o método de validação cruzada externa de *5-fold* utilizado para desenvolvimento dos modelos de QSAR (adaptado TROPSHA, 2010).

O conjunto de dados foi dividido em cinco subgrupos de tamanhos iguais, no qual um desses subgrupos (20% de todos os compostos) foi definido como conjunto de validação externa e os quatro conjuntos restantes formam o conjunto de modelagem (80% de todo o conjunto de dados). Esse procedimento foi repetido cinco vezes permitindo que cada um dos cinco subconjuntos seja usado como conjunto de validação externa. Os modelos foram gerados usando apenas o conjunto treinamento. É importante enfatizar que o conjunto de validação externa não pode ser empregado para construção e/ou seleção dos modelos. Cada conjunto de dados foi dividido em vários conjuntos de treinamento (utilizado para geração dos modelos de QSAR) e teste (utilizado para validação externa). Ao final, os modelos foram gerados usando compostos de cada conjunto treinamento e os conjuntos testes foram usados para avaliar a preditividade do modelo. A predição final é a média de todos os modelos.

Os modelos selecionados foram aplicados aos conjuntos de validação externa para predizer suas propriedades experimentais. Esse procedimento foi repetido cinco vezes para garantir que cada composto passasse uma vez pelo conjunto de validação externa. Como a taxa de acerto de cada modelo é estimada apenas nos compostos do conjunto de validação externa, os quais nunca são usados para gerar os modelos, esse protocolo garante uma estimação objetiva da real preditividade externa dos modelos. Além disso, 1.000 execuções de randomização da variável Y foram realizadas para garantir que a preditividade dos modelos gerados não foi devida a correlações aleatórias. Nesse procedimento, a variável Y é

aleatorizada e novos modelos são gerados. Caso a preditividade dos modelos com variável Y aleatória obtenha resultados melhores que os modelos de QSAR, então os modelos devem ser descartados, visto os descritores moleculares não descrevem bem a variável Y.

3.5 Parâmetros de avaliação dos modelos de QSAR de classificação

Existem dois erros de predição possíveis em um modelo binário (e.g., 0/1, sim/não, ativo/inativo): falso positivos e falso negativos. A performance de uma classificação binária é normalmente resumida em uma matriz de confusão (Tabela 3).

Tabela 3. Matriz de confusão de uma classificação binária.

	Condição positiva	Condição negativa
Predição positiva	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Predição negativa	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Tomando o valor 1 como classe positiva e 0 como classe negativa, verdadeiros positivos (VP) e verdadeiros negativos (VP). Um falso positivo (FP) ocorre quando o classificador estima 0 para uma classe, quando que na verdade ela pertence a classe 1. De maneira análoga, um falso negativo (FN) ocorre quando o classificador prediz 1 quando na verdade a classe é 0.

Os seguintes parâmetros foram utilizados para avaliar diferentes aspectos da performance dos modelos (Equações 5-12).

3.5.1 Acurácia (Acc) e Acurácia Balanceada (CCR)

Acurácia é a proporção de predições corretas (ambos verdadeiros positivos e verdadeiros negativos) entre o número total de casos examinados (Equação 5).

$$Acc = \frac{VP + VN}{P + N}$$

Quando uma das classes é muito maior que a outra (conjuntos desbalanceados), o modelo tende a aprender mais com a classe majoritária, o que pode desfavorecer a predição da classe minoritária. Para contornar esse problema, é utilizada a acurácia balanceada, que é a

média aritmética da sensibilidade e especificidade, ou a média da acurácia obtida para cada uma das classes (positiva e negativa). A acurácia balanceada, também chamada de taxa de classificação correta (CCR) é descrita pela Equação 6.

$$CCR = \frac{Sensibilidade + Especificidade}{2}$$

3.5.2 Sensibilidade e Especificidade

A sensibilidade mede a taxa de verdadeiros positivos que são corretamente classificados pelo modelo. A especificidade mede a proporção de verdadeiros negativos que são corretamente classificados. São definidas pelas Equações 7 e 8.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{7}$$

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{8}$$

3.5.3 Valor Preditivo Positivo (VPP) e Valor Preditivo Negativo (VPN)

Valor preditivo positivo é a porcentagem de amostras classificadas como positivas dado que sejam verdadeiramente positivas. Enquanto que valor preditivo negativo é a porcentagem de amostras classificadas como negativas dado que sejam verdadeiramente negativas. São definidas pelas seguintes equações (9 e 10):

$$VPP = \frac{VP}{VP + FP} \tag{9}$$

$$VPN = \frac{VN}{VN + FN} \tag{10}$$

3.5.4 Área sob a curva ROC (AUC)

Curva ROC (*receiver operating characteristic*) é uma representação gráfica que ilustra a robustez de um modelo binário em discriminar diferentes classes. A curva ROC é gerada a partir da taxa de verdadeiros positivos (sensibilidade) contra a taxa de falsos positivos (1 -

especificidade) em diferentes *thresholds*. A área sob curva ROC (AUC) representa a probabilidade de acerto de compostos ativos pelo modelo. Uma vantagem da AUC é que ela gera uma medida objetiva da qualidade do modelo, não dependo da quantidade de ativos e inativos nem da sua proporção, o que a torna uma propriedade do método, não do experimento.

3.5.5 Medida F (F1*score*)

A medida F (Equação 11) é uma avaliação da qualidade do modelo que considera precisão e sensibilidade (*recall*). Em casos de conjuntos desbalanceados em que a classe minoritária é mais importante, a medida F se mostra eficiente em avaliar a qualidade do modelo (POWERS, 2011).

$$F1 = 2 \cdot \frac{Precisão \cdot Se}{Precisão + Se}$$

3.5.6 Coeficiente Kappa de Cohen (Kappa)

Uma abordagem direta para normalizar a acurácia e reduzir a tendência é o Coeficiente Kappa de Cohen (Equação 12). O Kappa é uma medida mais robusta uma vez que leva em consideração a ocorrência por chance.

$$Kappa = \frac{P_{observado} - P_{esperado}}{1 - P_{esperado}}$$

3.6 Métodos de aprendizado de máquina

3.6.1 *Support Vector Machine (SVM)*

SVM é um método de aprendizado de máquina supervisionado para classificações binárias (reconhecimento de padrões) e valores reais de funções de aproximações (regressão) (MISHRA; TRIPATHI; YADAV, 2010). O SVM pertence à família dos modelos discriminantes; este método tenta encontrar uma combinação de amostras para construir um plano de maximização da margem de separação entre as classes (Figura 10).

É necessário o uso da função de *kernel* que possibilita a construção de um espaço de descritores *n*-dimensional. Várias funções de *kernel* são utilizadas, como funções lineares,

radiais, polinomial e sigmóides. Além disso, o desempenho de generalização do SVM dependem da definição de parâmetros do algoritmo (C e ε) (VAPNIK, 2000).

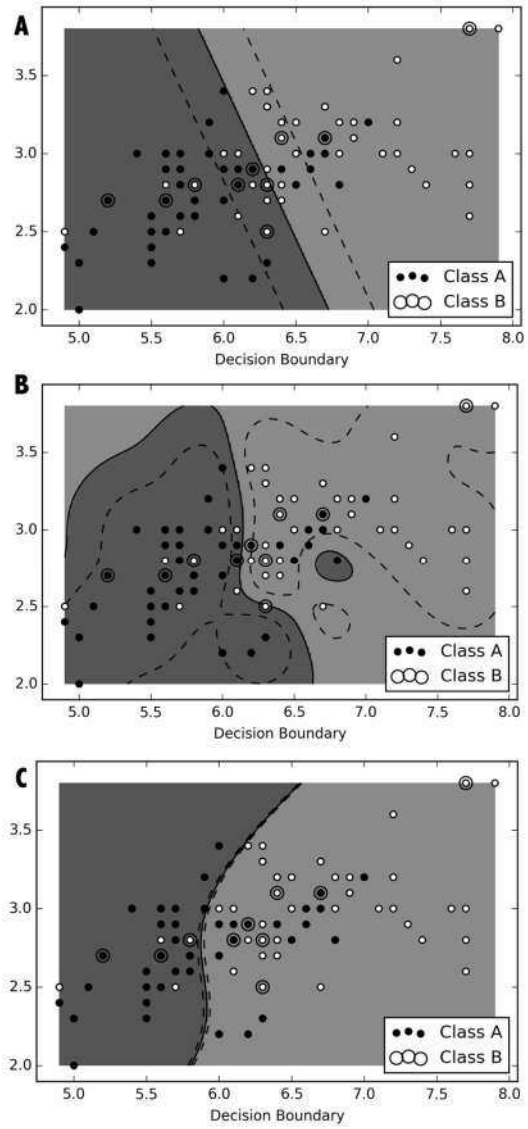

Figura 10. Exemplos de modelos de SVM desenvolvidos para classificar uma classe de dados multidimensionais de um conjunto da literatura (Iris Dataset). (A) SVM linear, (B) SVM radial e (C) SVM polinomial (Fonte: BRAGA et al., 2015).

3.6.2 *Gradient Boosting Machine* (GBM)

O algoritmo GBM gera modelos a partir de uma sequência de árvores de decisões, em que cada árvore é construída a partir dos resíduos de predição da árvore anterior. O particionamento dos dados é determinado em cada etapa algoritmo, promovendo o aumento das árvores. Em seguida, os desvios dos respectivos resíduos para cada partição de uma

árvore de decisão serão computados e posteriormente gerados na árvore seguinte. Dada à sequência anterior das árvores, o próximo nó da árvore será ajustado aos resíduos a partir de informações anteriores, a fim de encontrar outra partição que irá reduzir ainda mais a variância residual para os dados (NATEKIN; KNOLL, 2013, BERK, 2008).

3.6.3 *k- Nearest Neighbors (k-NN)*

k-NN é um método simples utilizado tanto para classificação ou regressão. Em ambos os casos, os vizinhos são definidos por uma função de distância, que busca a predição de um novo exemplo a partir de uma instância rotulada ou classe *k*. Em modelos de classificação, a saída é definida pela classe majoritária. Em modelos de regressão, o resultado é uma média dos resultados dos *k* vizinhos mais próximos (MITCHELL, 1997; CONSONNI; BALLABIO; TODESCHINI, 2009) (Figura 11).

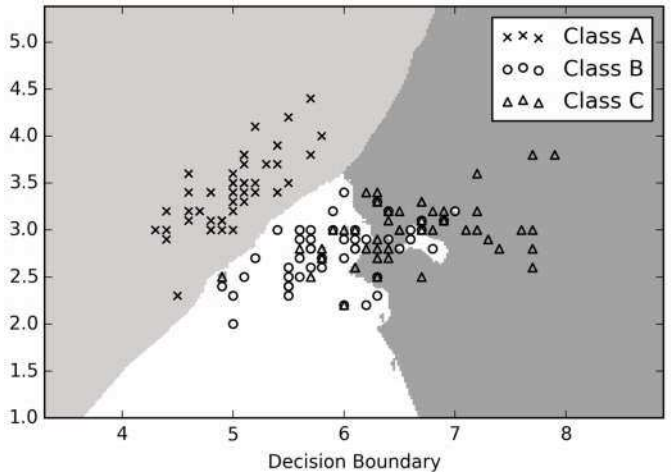

Figura 11. Exemplos de modelos de *k*-NN desenvolvidos para classificar uma classe de dados multidimensional de um conjunto da literatura (Iris Dataset). As três cores de fundo diferentes representam a fronteira de decisão (Fonte: BRAGA et al., 2015).

3.6.4 *Partial least squares discriminant analysis (PLS-DA)*

A análise discriminante por mínimos quadrados parciais (PLS-DA) é uma variação do método dos mínimos quadrados (PLS), utilizada quando a resposta Y (atividade biológica) é de natureza binária. O PLS-DA faz uma relação entre a matriz de covariância das variáveis independentes (descritores moleculares) e a atividade biológica (ZANETTI, 2014).

3.6.5 *Randon Forest (RF)*

Random Forest é um algoritmo que representa um conjunto de árvores de decisão individuais, no qual a saída de todas as árvores é agregada para se obter a predição final. Cada árvore é gerada por uma amostragem realizada por *bootstrap* do conjunto modelagem de *n*-compostos, formando uma árvore de decisão para o conjunto de treinamento. Os compostos que não estão no conjunto treinamento, são colocados no conjunto *out-of-bag* (OOB) (tamanho recomendado de ~ N / 3). Em seguida, a divisão é feita pelo algoritmo CART entre os *m* descritores selecionados aleatoriamente de todo o *pool* de descritores em cada nó. Cada árvore então cresce levando-se em consideração que os valores de classificação preditos são definidos por maioria dos votos para uma das classes. Assim, cada árvore prevê valores para apenas os compostos que não estão incluídos no conjunto de treinamento da árvore (para OOB definido apenas). Este método possui suas próprias características estatísticas (com base na predição conjunto OOB) que podem ser utilizados para validação e seleção destes modelos (BREIMAN et al., 1984) (Figura 12).

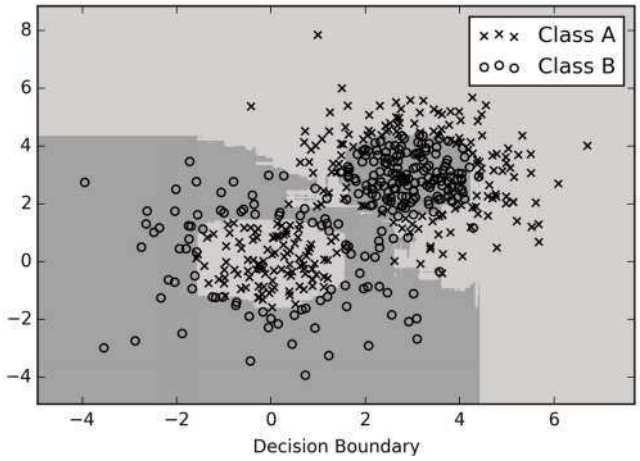

Figura 12. Exemplo de classificação por RF desenvolvidos para classificar uma classe de dados multidimensional de um conjunto da literatura (Iris Dataset). As duas cores de fundo diferentes representam a fronteira de decisão usada para o classificador (Fonte: BRAGA et al., 2015).

4 RESULTADOS E DISCUSSÃO

4.1 Substratos de CYP3A4

4.1.1 Caracterização do conjunto de dados

4.1.1.1 Conjunto de dados de substratos de CYP3A4

O conjunto de dados de substratos de CYP3A4 utilizado neste trabalho consistiu em 8.214 compostos, dos quais 475 são substratos de CYP3A4 (ZARETZKI et al., 2012) e 7.739 são não-substratos de CYP3A4, que foram retirados do banco de dados PubChem bioassay (ID:1851) e representam o maior conjunto de dados de substratos de CYP3A4 disponível de domínio público.O processo de preparo do conjunto de dados foi realizado utilizando o protocolo *KSAR*, que incluiu vários módulos para o preparo dos dados, como a padronização das estruturas e remoção de duplicatas, como descrito previamente na Figura 7.

4.2 Geração dos modelos de QSAR

Para geração dos modelos de QSAR, utilizou-se um fluxo de trabalho totalmente integrado aos programas R e KNIME, que inclui vários módulos, tais como processamento de dados (ex. remoção de duplicatas), módulo de divisão racional do conjunto de dados (ex. algoritmo de *Kennard-Stone*) e módulo de divisão de dados aleatório. A utilização deste fluxo de trabalho no KNIME permite a integração de vários pacotes computacionais, como o *Stantardizer*, *Isida duplicates*, *HitQSAR*, entre outros, a fim de executar diversas etapas da modelagem de forma interativa. Também foram incluídos vários métodos de aprendizado de máquina (AM), o que possibilitou a geração de modelos com diferentes especificidades no reconhecimento de padrões.

O procedimento de validação cruzada *5-fold* foi utilizado para estimar a robustez dos modelos utilizando o conjunto treinamento, enquanto que o conjunto teste foi aplicado para validar e estimar o poder preditivo dos modelos gerados.

Para isso, foram realizadas seis divisões de conjuntos de dados com seis proporções diferentes de balanceamento, utilizando a técnica de *Undersampling*. Esta técnica se baseia na modificação da disposição dos dados em relação à distribuição das classes desbalanceadas (EITRICH et al., 2007) permitindo equilibrar os dados de acordo com a disposição das classes minoritárias em relação à proporção das classes majoritárias. As classes minoritárias são selecionadas aleatoriamente, para coincidir com o número de dados contidos na classe

majoritária. Utilizou-se o algoritmo *k*-NN, que busca averiguar a distância de uma classe em relação à outra classe, induzido uma distribuição favorável dos dados.

O primeiro conjunto de dados foi gerado utilizando o conjunto de dados total disponível na literatura e no PubChem, que portanto está distribuído de maneira desbalanceada, 475 substratos e 7.739 não-substratos (conjunto A). O segundo conjunto de dados (conjunto B, proporção 1:1 – conjunto balanceado), foi constituído por 475 substratos e 475 não-substratos. O terceiro conjunto de dados, de proporção 1:2 (conjunto C), foi constituído por 475 substratos e 950 não-substratos. O quarto conjunto de dados, de proporção 1:3 (conjunto D) possui 475 substratos e 1.425 não-substratos. O quinto conjunto de dados, de proporção 1:4 (conjunto E), possui 475 substratos e 1.900 não-substratos, e por último o sexto conjunto (conjunto F) possui 475 substratos e 3.870, resultando em uma proporção de 1:8.

Foram gerados vários modelos de QSAR utilizando os seis conjuntos de dados (A-F) com diferentes proporções de balanceamento de substratos e não-substratos. Para cada conjunto, foram gerados vários modelos utilizando descritor MACCS e variando os algoritmos de AM. Os resultados estatísticos dos melhores modelos gerados para os seis conjuntos de dados estão apresentados nas Figuras 13 e 14 para o conjunto treinamento e teste, respectivamente. Os resultados estatísticos completos dessa avaliação estão dispostos nas Tabelas 5-16 disponíveis no Apêndice I.

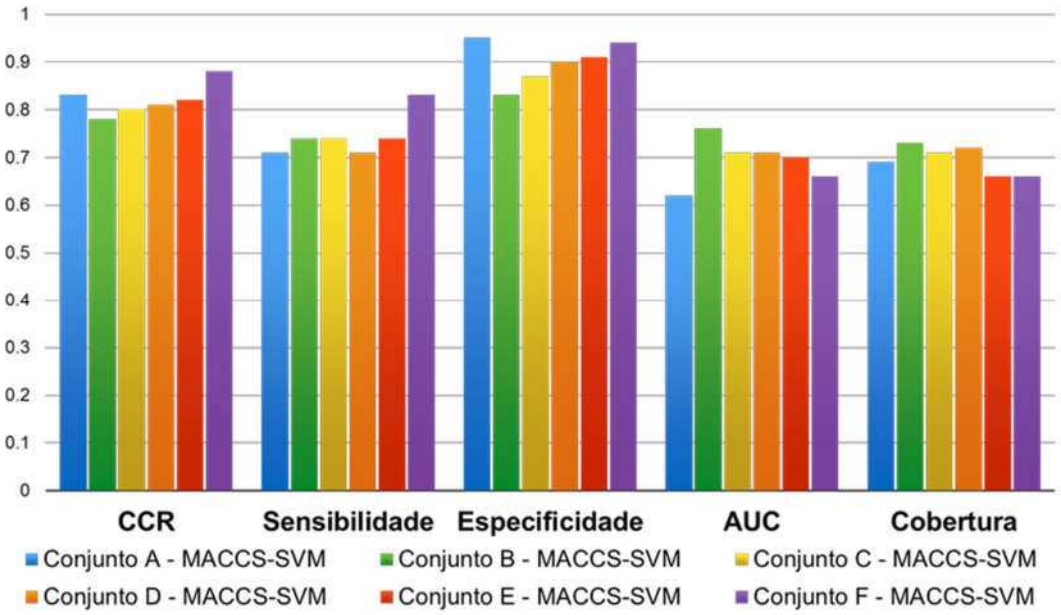

Figura 13. Características estatísticas dos melhores modelos de QSAR gerados para substratos CYP3A4 avaliados por *5-fold* do conjunto treinamento utilizando os seis conjuntos de dados. CCR: taxa de classificação correta; AUC: área sob a curva ROC.

A combinação do descritor MACCS com diferentes métodos de aprendizado de máquina gerou modelos robustos e preditivos, com valores de acurácia balanceada (CCR) variando entre 0,77 e 0,88 e cobertura entre 0,53 e 0,73 (Tabelas 5-16) Apêndice I.

O primeiro parâmetro que deve ser avaliado é a taxa de classificação correta (CCR) ou acurácia balanceada, que é calculada a partir da média aritmética entre a sensibilidade e especificidade, e reflete na proporção de predições corretas, atribuindo o mesmo peso para as duas classes (positivo e negativo). Em seguida, é importante observar a sensibilidade e especificidade, que não devem ser muito diferentes uma da outra para o mesmo modelo, ou seja, o modelo deve ser capaz de corretamente classificar os verdadeiros positivos na mesma proporção que classifica corretamente os verdadeiros negativos. A AUC é uma medida objetiva da performance do classificador, e não depende da quantidade de ativos e inativos nem da sua proporção, sendo também muito importante a sua avaliação. Já a cobertura é a porcentagem de compostos preditos que estão dentro do DA. Valores iguais a 1 para a cobertura podem ser obtidos para modelos que não estimaram o DA, ou quando 100% dos compostos estão dentro do DA.

Como pode ser observado na Figura 13, o conjunto F apresentou os melhores valores de CCR, sensibilidade, seguido pelo conjunto A (desbalanceado). No entanto, avaliando-se a AUC e a cobertura, o conjunto B, que é o conjunto balanceado, foi superior para avaliação do conjunto treinamento.

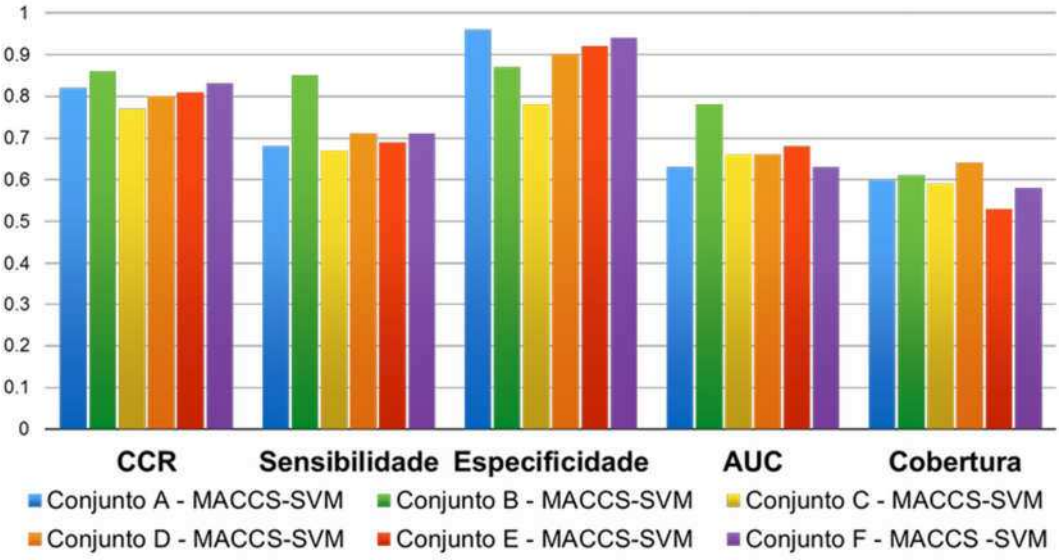

Figura 14. Características estatísticas dos melhores modelos de QSAR para substratos CYP3A4 avaliados para o conjunto teste utilizando os seis conjuntos de dados. CCR: taxa de classificação correta; AUC: área sob a curva ROC.

Como pode ser observado na Figura 14, que apresenta os resultados do conjunto teste, o conjunto de dados B (1:1, balanceado) apresentou melhores valores de CCR, sensibilidade e AUC do que todos os outros modelos, além de apresentar valores de sensibilidade e especificidade próximos um do outro. No entanto, ao observamos a performance do modelo gerado para o conjunto A (totalmente desbalanceado) para o conjunto teste (Figura 13), é possível verificar que este modelo apresentou a menor sensibilidade, além da maior discrepância entre os valores de sensibilidade e especificidade, o que mostra que o modelo consegue classificar melhor os compostos da classe inativa (não-substratos), visto que esta é a maior classe. E ainda nesse sentido, observamos que este modelo apresentou menor valor de AUC, juntamente com o modelo gerado para o conjunto F (1:8). Assim, o conjunto de dados B (balanceado 1:1) foi selecionado como o melhor balanceamento do conjunto de dados para a geração de modelos de QSAR de classificação para os substratos de CYP3A4, e foi utilizado na otimização dos modelos.

A seleção do descritor MACCS para geração dos modelos iniciais justifica-se, pois este descritor fornece uma e representação computacional das estruturas a partir de uma coleção de 166 subestruturas predefinidas. Isto contribui para a realização de estudos de comparação, como similaridade estrutural. Dessa forma, os descritores de natureza de impressão digital molecular (*fingerprint)*, como o descritor MACCS, são mais adequados para esta etapa inicial do estudo, que objetivou a seleção do melhor balanceamento para a geração de modelos de QSAR.

4.2.1 Otimização de modelos de QSAR para substratos de CYP3A4

Dentre os seis tipos de balanceamento testados, o conjunto de dados balanceado (Conjunto B, proporção 1:1) foi selecionado para o prosseguimento do estudo, por apresentar os melhores valores estatísticos para o conjunto treinamento e conjunto teste. Realizou-se a otimização dos modelos de classificação utilizando diferentes tipos de descritores e diferentes algoritmos de classificação. Essa etapa teve como finalidade de melhor explorar as informações estruturais através da variação dos tipos de descritores, para classificar corretamente os compostos em substratos e não-substratos de CYP3A4. Os modelos foram gerados variando quatro tipos de descritores moleculares (*AtomPair*, *PubChem*, MACCS e *FeatMorgan*) e quatro métodos de AM (SVM, GBM, PLSDA e *k*-NN), resultando ao total em 16 modelos de QSAR. Os resultados estatísticos dessa análise para os melhores modelos de

QSAR gerados para conjunto de dados balanceado (Conjunto B) podem ser observados nas Figuras 14 e 15 para os conjuntos treinamento e teste, respectivamente. As tabelas 17, 18 e 19 no Apêndice I apresentam os resultados estatísticos detalhados para os 16 modelos gerados.

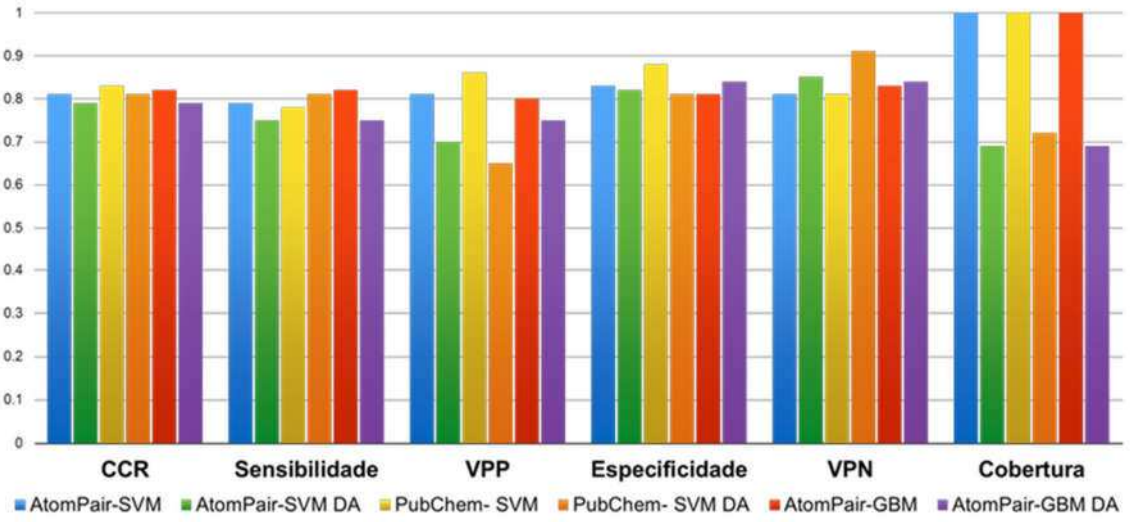

Figura 15. Características estatísticas dos modelos de QSAR para substratos de CYP3A4 avaliados por *5-fold* para o conjunto treinamento. DA: Domínio de aplicabilidade; CCR: taxa de classificação correta; VPP: valor preditivo positivo; VPN: valor preditivo negativo.

A combinação de diferentes tipos de descritores e de métodos de aprendizado de máquina levaram a geração de modelos QSAR mais robustos e preditivos, com taxa de classificação correta (CCR) que variam entre 0,65-0,83 e cobertura de 0,69-0,89. Nas Tabelas 14 e 15, são apresentados apenas os três melhores modelos (AtomPair-SVM, PubChem-SVM e AtomPair-GBM), considerando ou não do DA.

É possível observar que os modelos gerados a partir dessas três combinações de descritores e métodos de AM apresentaram valores estatísticos muito próximos (Figura 15). O modelo PubChem-SVM sem considerar DA apresentou resultados ligeiramente superiores de CCR, VPP, VPN, especificidade e cobertura, dentre os outros modelos. O VPP reflete a probabilidade de uma amostra ser classificada como positiva caso realmente ela seja positiva, enquanto que o VPN é a probabilidade de uma amostra ser classificada como negativa caso realmente ela seja negativa. No entanto, o modelo gerado pela combinação AtomPair-GBM sem considerar DA apresentou maior sensibilidade e menor diferença entre os valores de sensibilidade e especificidade, o que torna este modelo mais adequado para classificar corretamente tanto os substratos quanto os não substratos de CYP3A4.

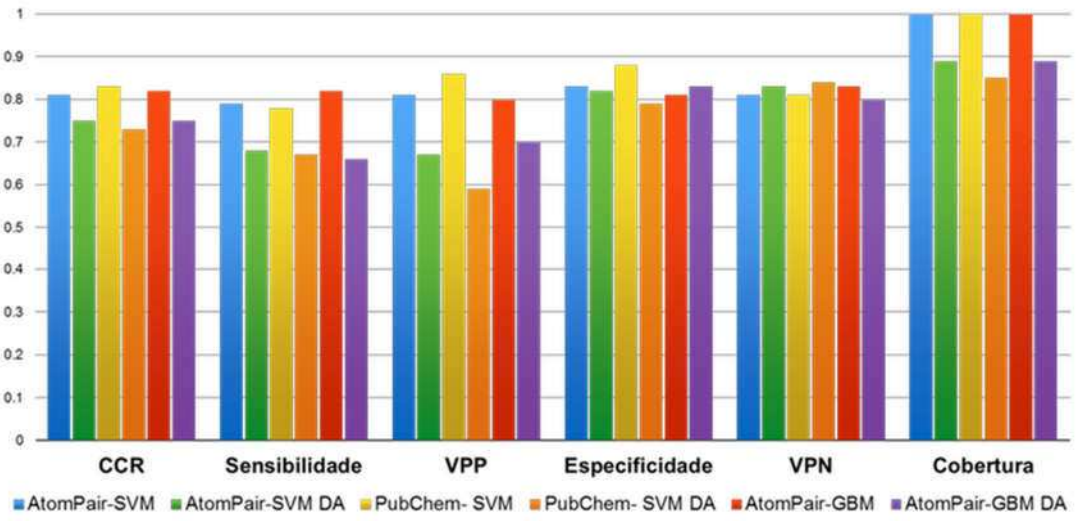

Figura 16. Características estatísticas dos modelos de QSAR para substratos de CYP3A4 avaliados para o conjunto teste. DA: Domínio de aplicabilidade; CRCR: taxa de classificação correta; VPP: valor preditivo positivo; VPN: valor preditivo negativo.

Ao analisar os resultados estatísticos dos modelos gerados utilizando o conjunto teste (Figura 16), mais uma vez observa-se que estes modelos são muito semelhantes. No entanto, observa-se que o modelo AtomPair-GBM apresentou a menor diferença entre os valores de sensibilidade e especificidade, e ainda uma maior cobertura. Observou-se ainda que alguns modelos obtiveram resultados discrepantes ao considerar ou não o DA, sendo que alguns destes modelos apresentavam os melhores valores estatísticos sem considerar o DA, como foi o caso do AtomPair-SVM.

A fim de verificar se a preditividade dos modelos gerados foi ou não devida a correlações aleatórias, realizou-se o teste da randomização da variável Y. Os resultados estão reportados na Tabela 4.

Tabela 4. Resultados estatísticos para os melhores modelos de QSAR para substratos de CYP-3A4 avaliados pela randomização da variável Y.

Modelo	Acurácia	CCR	Kappa	Se	Sp	AUC
AtomPair-SVM	0,53±0,03	0,53±0,12	0,05±0,06	0,30±0,13	0,75±0,11	0,54±0,04
PubChem-SVM	0,53±0,03	0,52±0,15	0,04±0,06	0,29±0,15	0,75±0,15	0,54±0,03
AtomPair–GBM	0,54±0,04	0,32±0,06	0,09±0,08	0,51±0,07	0,57±0,05	0,54±0,03

SVM: *Support Vector Machine*; GBM: *Gradient boosting Method:* CCR: Taxa de classificação correta; Kappa: coeficiente Kappa de Cohen; Se: Sensibilidade; Sp: Especificidade; AUC: Área sob a curva ROC.

Como pode ser observado na Tabela 4, os resultados estatísticos apresentados para os três modelos de QSAR estão à maioria em torno de 0,50, o que significa que estão na faixa de uma predição randômica. Observam-se ainda alguns valores mais baixos que 0,50, o que significa que esses modelos gerados pela randomização da variável Y não são aceitáveis. Esses resultados demonstram que e as análises não são aleatórias, ou seja, que os três melhores modelos de QSAR são altamente preditivos, consistentes e estatisticamente significantes para identificar substratos de CYP3A4.

4.3 Inibidores de CYP3A4

4.3.1 Caracterização do conjunto de dados

4.3.1.1 Conjunto de dados de inibidores de CYP3A4

O conjunto de dados de inibidores de CYP3A4 utilizado consiste em um dos maiores conjuntos de dados disponíveis publicamente para CYP3A4. O conjunto de dados possui 42.294 compostos com vários tipos de ensaios para esta enzima. Contudo, durante o extensivo processo de preparo dos dados, vários aspectos merecem atenção especial com o objetivo de padronizar o conjunto de dados.

Durante o preparo do conjunto de dados, primeiramente realizou-se a mineração dos dados (*data mining*). Os 42.294 compostos com vários tipos de atividade em CYP3A4 foram extraídos automaticamente do ChEMBL, e as informações das atividades dos compostos foram agrupadas. Foram selecionados os compostos com propriedade*s* específicas (IC_{50}; *K*i; EC_{50}), ou seja, com as propriedades biológicas para o estudo, compostos que não se encaixaram na propriedade definida foram removidos, restando nesta etapa o total de 21.710 compostos. Em seguida, uma regra foi gerada para filtrar os compostos em indutores, substratos e inconclusivos. Os compostos encontrados foram removidos, restando um total de 9.368 compostos inibidores de CYP3A4. A próxima etapa do preparo do conjunto de dados consistiu na inspeção manual. Os 9.368 compostos foram conferidos manualmente com a finalidade de verificar se ainda havia indutores e substratos remanescentes. Um total de 10% dos compostos (938) foi verificado manualmente com a literatura, a fim de garantir uma boa confiabilidade dos dados. Nesta análise não foi encontrado nenhum substrato ou indutor remanescente. Dessa forma, o conjunto de dados permaneceu com 9.368 compostos.

Ao final do preparo dos dados, foi definido o limiar de atividade (*threshold)* para classificação de inibidores e não inibidores. Essa definição não está clara na literatura. Na

literatura, observou-se um limiar de atividade < 10μM para considerar uma interação com CYP3A4 (MAYHEW; JONES; HALL, 2000; KHOJASTEH et al., 2011; UENG et al., 2013). Por isso, definiu-se neste trabalho que compostos com valores de propriedade < 10 μM foram considerados como inibidores (4.962) e compostos com valores de atividade > 10μM foram considerados como não inibidores (4.224). Para construção de modelos multiclasse, o limiar de atividade foi definido como se segue: inibidor forte ≤ 1 μM; inibidor fraco-moderado, propriedade entre 1 μM e 10 μM; não inibidor ≥ 10 μM. Nesta análise, 182 compostos foram considerados como inconclusivos, pois apresentavam valores de atividade que poderiam cair em ambos os limiares de atividade (inibidor e não inibidor). Por exemplo, valores < 20 μM, poderiam ser maiores ou menores que 10 μM, e serem considerados como tanto inibidor e não inibidor. Esses 182 compostos foram excluídos do conjunto de dados, restando no total 9.137 compostos.

As estruturas químicas foram devidamente padronizadas, foram calculados os tautômeros, realizou-se a otimização da estrutura 2D, removeu-se os contraíons e normalizou-se os quimiotipos específicos como grupos aromáticos e nitro, sais inorgânicos, compostos organometálicos, polímeros e misturas também foram removidos.

A existência de duplicatas, ou seja, compostos idênticos reportados mais de uma vez no mesmo conjunto de dados, é conhecida por levar a uma sobrestimação da preditividade dos modelos de QSAR. Neste processo, foram removidas 1.756 duplicatas, destas 193 eram duplicatas discordantes, ou seja, os dados de atividade para os mesmos compostos eram diferentes, existindo um grande desvio dos valores experimentais, indicando assim, que a qualidade desses dados é baixa. As 193 duplicatas foram verificadas manualmente em nossos registros. Quando os valores experimentais associados a um mesmo composto eram idênticos, apenas um dos registros foi mantido no conjunto de dados. Caso contrário, todos os registros foram removidos. Os programas HiTQSAR e *ISIDA Duplicates* foram utilizados para conferir novamente as duplicatas para analisar exaustivamente se ainda restou alguma duplicata remanescente.Vale ressaltar que houve 89% de concordância entre os dados do conjunto utilizado, o que garante uma boa qualidade dos dados extraídos do banco de dados CheMBL. Ao final de todo o preparo e padronização, o conjunto de dados de inibidores de CYP3A4 totalizou em 7.684 compostos para a realização da modelagem de QSAR.

4.3.2 Geração dos modelos de QSAR para inibidores de CYP3A4

Os modelos para inibidores de CYP3A4 foram gerados utilizando apenas um tipo de descritor molecular (*Morgan*) e dois métodos de AM (SVM e RF), resultando ao total em 4 modelos, sendo 2 binários e 2 multiclasse. Os resultados estatísticos para os modelos de classificação (binários) e modelos multiclasse (semi-quantitativos) gerados podem ser observados na Figura 17. Os resultados completos dessa análise estão reportados na Tabela 20 no apêndice I.

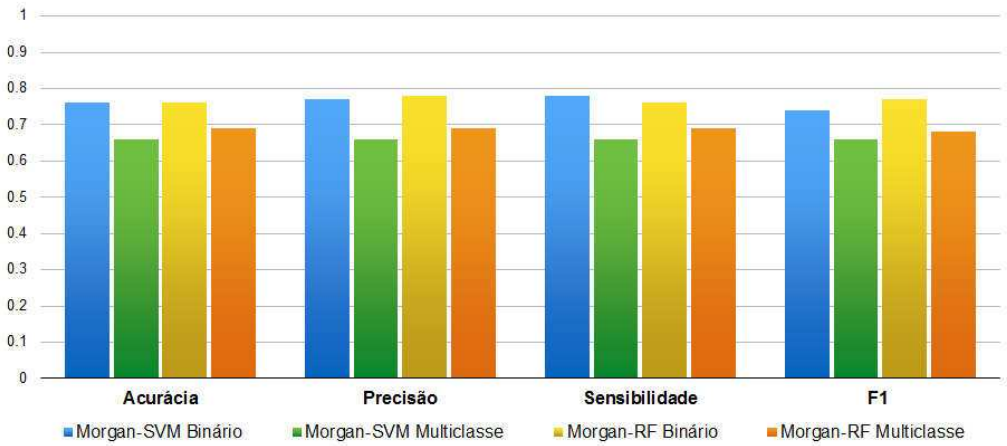

Figura 17. Resultados estatísticos dos melhores modelos de QSAR binário e multiclasse para inibidores de CYP3A4 avaliados por *5-fold*.

Para geração dos modelos de QSAR binário e multiclasse para inibidores de CYP3A4, utilizou-se a técnica de 5-*fold*, ou seja, dividiu-se o conjunto de dados em conjunto modelagem e conjunto de validação externa, não sendo utilizado um conjunto adicional para validação externa. Os resultados reportados na Figura 17 representam a predição final da média de todos os modelos gerados.

Para possibilitar uma comparação entre os resultados estatísticos para os modelos binários e multiclasse para inibidores de CYP3A4, utilizou-se apenas os parâmetros que fossem possíveis para os dois tipos de modelos (binário e multiclasse). Nesse caso, os parâmetros estatísticos avaliados foram acurácia, precisão, sensibilidade e valor de F1.

Como pode ser observado na Figura 17, os dois melhores modelos binários obtidos foram Morgan-SVM e Morgan-RF. Estes modelos apresentaram valores iguais de acurácia de 0,76, que corresponde à porcentagem de moléculas que são classificadas corretamente pelo modelo. Além disso, apresentaram valores de sensibilidade de 0,74 e 0,77, respectivamente. A sensibilidade avalia a proporção de casos reais de positivos (inibidores) que são corretamente

preditos como positivos. A precisão destes modelos foi de 0,77 e 0,78, respectivamente, enquanto que o F1 foi de 0,76 e para ambos os modelos. Estes dois modelos binários se apresentaram muito semelhantes quanto aos parâmetros estatísticos.

A geração de modelos multiclasse tem por finalidade o reconhecimento de padrões e classes dentro do conjunto de dados utilizado na geração dos modelos. Diferentemente dos modelos binários, o reconhecimento de padrões ou a geração do classificador, serão atribuídas por mais de duas classes. Portanto, modelos multiclasse são capazes de reconhecer mais classes que os modelos binários, ou seja, podem classificar a intensidade da propriedade.

Como pode ser observado na Figura 17, os dois melhores modelos multiclasse também foram gerados utilizando a combinação de Morgan-SVM e Morgan-RF. O modelo Morgan-RF apresentou um valor de precisão de 0,69, enquanto que o modelo Morgan-SVM, 0,66. O modelo Morgan-RF também se mostrou ligeiramente superior com relação ao valor de F1, com valor de 0,69, comparado ao valor de 0,66 para o modelo Morgan-SVM. No entanto, diante desses resultados, os modelos multiclasse Morgan-SVM e Morgan-RF foram também muito semelhantes, sendo ambos considerados os melhores modelos para avaliação da inibição de CYP3A4.

A utilização dos modelos binários e multiclasse buscam resolver problemáticas encontradas na distribuição de classes presentes nos conjunto de dados, como também na escolha de modelos mais robustos para classificar uma determinada propriedade. Entretanto, os modelos multiclasse buscam classificar as informações contidas no modelo em 3 ou mais classes, podendo ser exemplificado no uso de um modelo no qual classifica inibidores para qual tipo de isoforma, ou a intensidade de um inibidor - fraco, moderado e forte - para uma determinada isoforma de CYP.

Os modelos de QSAR obtidos neste trabalho podem ser muito úteis na predição do perfil metabólico, considerando a enzima CYP3A4 para novos tipos de possíveis fármacos, pois, a identificação de inibidores de CYP3A4 de novas entidades químicas pode auxiliar na análise de interação fármaco-fármaco. Certamente, este estudo preditivo reduzirá drasticamente o tempo e o impacto de custos no planejamento e desenvolvimento de novos fármacos.

4.3.3 Mapas de Probabilidade Predita (MPPs)

Os resultados obtidos pelos modelos finais de QSAR são de difícil compreensão mecanística. Contudo, na busca por ferramentas que auxiliam na interpretação desses modelos, foram compilados mapas que permitissem visualizar fragmentos estruturais favoráveis (positivo) e desfavoráveis (negativo) para a inibição de CYP3A4. Esses mapas são conhecidos como MPPs, e se caracterizam pela formação de circunferências radiais que distinguirão os fragmentos favoráveis ou desfavoráveis na estrutura da molécula. Nos fragmentos que apresentam contribuição positiva para inibição da CYP3A4, o raio ao redor do fragmento estrutural será delineado em cor verde. Já um fragmento que está contornado por uma circunferência radial rosa terá a propriedade desfavorável para a inibição da enzima. No entanto, existem grupos ou átomos delineados em cor cinza, que representam grupos sem nenhuma contrição para atividade inibitória de CYP3A4. Vale ressaltar ainda, que a intensidade da cor destes círculos indica a intensidade da resposta de acordo com a cor apresentada. Uma vez que os fragmentos que apresentam maior intensidade na cor e maior intensidade do contorno, terão uma maior contribuição para a propriedade.

Selecionou-se alguns compostos do conjunto de dados para a visualização dos MPPs, mas essa análise é possível para todo e qualquer composto do conjunto de dados ou ainda para compostos que não fizeram parte da geração dos modelos.

Os mapas, apresentados na Figura 18 apresentam os fármacos abacavir, lapatinibe e ritonavir.

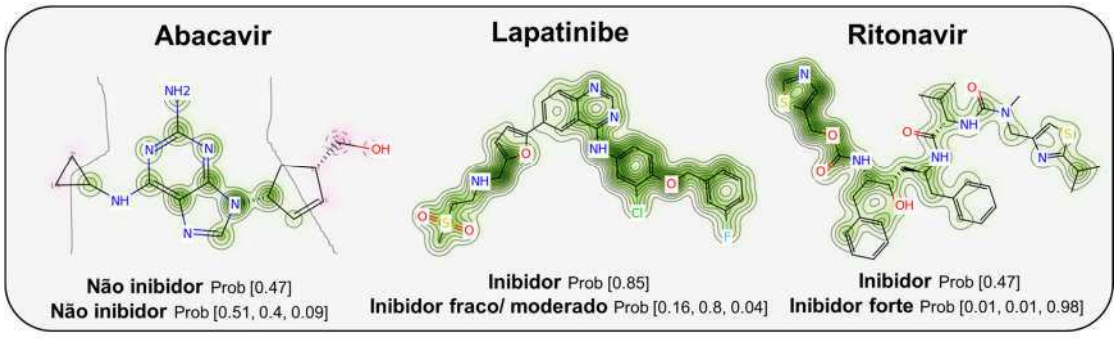

Figura 18. Mapa de probabilidade predita (MPP) para o modelo Morgan-RF de modelos binários de inibição de CYP34A. Bits verdes (átomos/fragmentos): contribuição favorável na propriedade (inibidor da CYP3A4); Bits rosa: contribuição desfavorável na propriedade (não inibidor de CYP3A4); Bits cinza: contribuição neutra na propriedade. Linhas em cinza delimitam a separação da contribuição desfavorável e favorável. O tamanho do vetor-bit de Morgan foi 1024 bits.

O abacavir é um antirretroviral usado em pacientes com HIV, é um substrato de CYP3A4 e não é inibidor da enzima (MANUSCRIPT et al., 2010). De acordo com a estrutura química do fármaco, pode-se observar que regiões delineadas com maior intensidade em verde são o anel pirimidina e o imidazólico. Esses fragmentos possuem características favoráveis para o reconhecimento deste fármaco como não inibidor de CYP3A4. Além disso, as linhas em cinza demarcam a separação entre os fragmentos que apresentam contribuição positiva e negativa na propriedade. A hidroxila delineada em rosa tem contribuição desfavorável para a propriedade de inibição. O abacavir foi classificado pelo modelo binário como não inibidor de CYP3A4, e pelo modelo multiclasse como não inibidor com probabilidade de 0,51% ficando na região de fronteira com o modelo binário, que foi 0,47%, neste caso e necessário realizar uma avaliação da potencialidade de inibição na enzima.

O lapatinibe é um fármaco antitumoral inibidor de CYP3A4 (HO et al., 2015), foi classificado pelo modelo binário como inibidor, já com relação ao modelo multiclasse foi classificado como inibidor fraco/moderado com uma probabilidade de 0,80%. Ao analisar a estrutura do fármaco, pode-se observar que as regiões delineadas com maior intensidade em verde, são as aminas secundárias próximas ao anel furano, e a pirimidina. Nesta estrutura é possível visualizar que as maiorias dos fragmentos presentes na estrutura apresentam contribuição favorável para a propriedade de inibição da enzima.

O ritonavir é um fármaco antirretroviral (MANUSCRIPT et al., 2010) e foi classificado pelo modelo binário como inibidor de CYP3A4, e pelo modelo multiclasse foi classificado como inibidor forte com probabilidade de 0,98. Os fragmentos presentes na estrutura do ritonavir indicam que o anel tiazol, que esta ao lado do oxigênio, está contornado com maior intensidade (bits verde), demonstrando assim, que a região apresenta contribuição favorável na propriedade inibitória. Os outros fragmentos também apresentam contribuição favorável, porem com menor intensidade no contorno.

Os mapas, apresentados na Figura 19 apresentam os fármacos cetoconazol, tioconazol e miconazol.

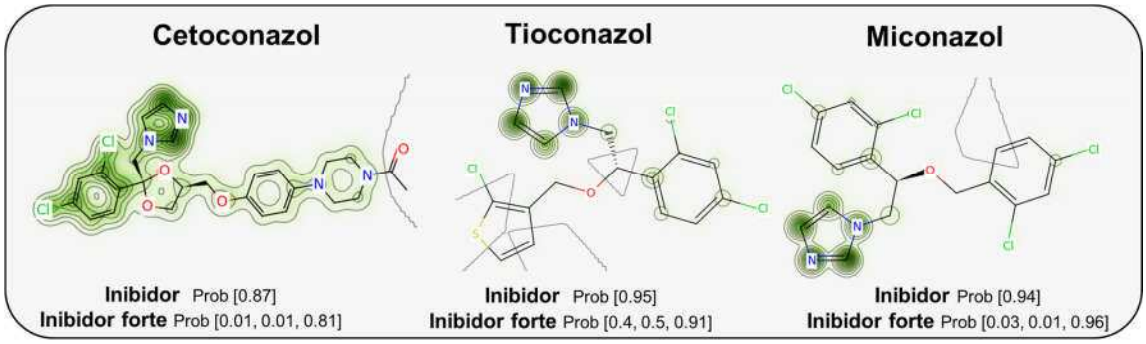

Figura 19. Mapa de probabilidade predita (MPP) para o modelo Morgan-RF de modelos binários de inibição de CYP34A. Bits verdes (átomos/fragmentos): contribuição favorável na propriedade (inibidor da CYP3A4); Bits rosa: contribuição desfavorável na propriedade (não inibidor de CYP3A4); Bits cinza: contribuição neutra na propriedade. Linhas em cinza delimitam a separação da contribuição desfavorável e favorável. O tamanho do vetor-bit de Morgan foi 1024 bits.

O cetoconazol é utilizado como controle positivo em muitos ensaios enzimáticos de inibição para CYP3A4, é considerado inibidor forte da enzima (DOSHI; LI, 2011). Como pode ser observado no mapa, o fragmento imidazol apresenta um átomo como nitrogênio que é responsável pelo reconhecimento do sítio ativo da CYP3A4. Ao observar a estrutura do cetoconazol, nota-se que o fragmento imadazólico esta delineado em verde com muita intensidade, indicando que a região apresenta contribuição favorável para reconhecer o fármaco como inibidor de CYP3A4, ou seja, quanto maior a intensidade do bit, maior a contribuição para classificar o composto em inibidor ou não inibidor. Fato importante a ser ressaltado é que grande maioria destes fármacos apresentam contribuição positiva em fragmentos estruturais nitrogenados. Isso ocorre principalmente pelas características que o nitrogênio apresenta de se coordenar com os átomos de ferro do grupo heme. As linhas em cinza delimitam a separação do grupo cetona ao restante da estrutura, indicando que o fragmento apresenta contribuição neutra, o que pode ser observado pela intensidade do bit, que está em cinza claro.

O miconazol e o tioconazol são fármacos com atividade antifúngica, são considerados inibidores de CYP3A4 (FERNÁNDEZ et al., 2009). Os dois fármacos foram classificados pelo modelo binário como inibidor da enzima, e pelo modelo multiclasse foram considerados inibidores forte com alta probabilidade. O fragmento imidazol em suas estruturas, delineado em verde indicam que esse fragmento possui características favoráveis para propriedade de inibição, como foi discutido anteriormente, esses fragmentos possuem átomos que são

capazes de se coordenar com o ferro do grupo heme. Os anéis fenila e tiofeno estão delineados em coloração cinza, o que caracteriza contribuição neutra para a propriedade. E as linhas em cinza demarcam as regiões que tem contribuição favorável e desfavorável.

Os mapas dos fármacos diltiazem, claritromicina e nefazodona estão representados na Figura 20.

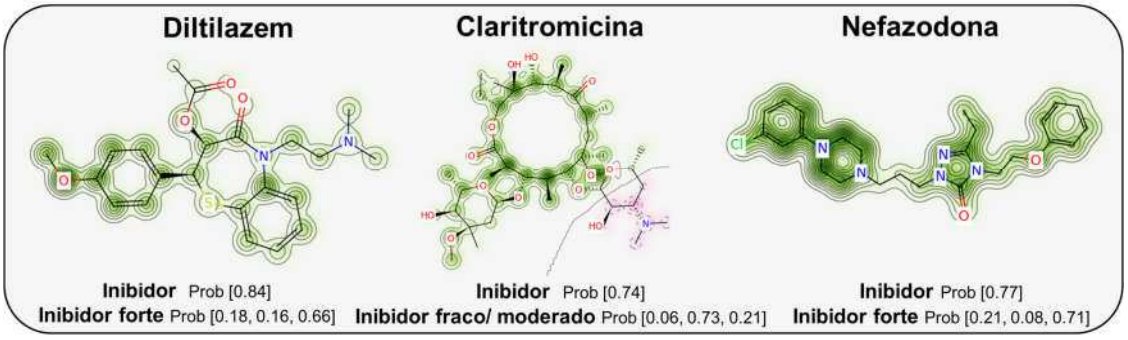

Figura 20. Mapa de probabilidade predita (MPP) para o modelo Morgan-RF de modelos binários de inibição de CYP34A. Bits verdes (átomos/fragmentos): contribuição favorável na propriedade (inibidor da CYP3A4); Bits rosa: contribuição desfavorável na propriedade (não inibidor de CYP3A4); Bits cinza: contribuição neutra na propriedade. Linhas em cinza delimitam a separação da contribuição desfavorável e favorável. O tamanho do vetor-bit de Morgan foi 1024 bits.

O diltiazem é um fármaco anti-hipertensivo, é considerado inibidor de CYP3A4 (SUTTON et al., 1997; FOTI et al., 2012). Foi classificado pelo modelo binário como inibidor da enzima, e pelo modelo multiclasse foi classificado como inibidor forte. Os fragmentos delineados no diltiazem em cor verde indicam que esses fragmentos possuem características favoráveis para a propriedade. No entanto, o anel aromático ligado ao éter, está delineado com maior intensidade, o que caracteriza uma maior contribuição para a propriedade de inibição de CYP3A4.

A claritromicina é um antibiótico da classe dos macrolídeos. É inibidor de CYP3A4 (LI et al., 2014). Pelo modelo binário foi classificado como inibidor, e pelo modelo multiclasse foi considerado inibidor fraco-moderado com probabilidade de 0,73. Como pode ser observado na estrutura do fármaco, o anel lactona macrolídeo da claritromicina e a cladinose estão delineados em cor verde. Esses fragmentos apresentam contribuição favorável para propriedade de inibição. Linhas em cinza delimitam as regiões que apresentam contribuição favorável e desfavorável. O grupo desosamina foi separado do restante da

estrutura, e está delineado em cor rosa que indica que o fragmento apresenta contribuição desfavorável para a propriedade de inibição de CYP3A4.

A nefazodona é um fármaco com atividade antidepressiva, que foi proscrito do mercado em alguns países da Europa e no Canadá, devido ao aparecimento de lesões hepáticas após o uso deste medicamento por alguns pacientes. Além da presença de toxicidade hepática, este fármaco apresenta uma alta atividade inibitória para a CYP3A4 (FERNÁNDEZ et al., 2009). Foi classificada pelo modelo binário como inibidor de CYP3A4, e pelo modelo multiclasse como inibidor forte com probabilidade de 0,71%. De acordo com os mapas de contornos gerados, os fragmentos delineados em verde com maior intensidade presentes na estrutura da nefazodona tem contribuição favorável para a propriedade de inibição da CYP3A4 são eles: o anel triazolona, piperazina e os 2 anéis fenila. Os fragmentos de triazolona e piperazina apresentam átomos capazes de se coordenar com o ferro, que contribuem para interação com o Fe^{3+} do grupamento heme.

Os mapas, apresentados na Figura 21 apresentam os fármacos nicardipina, alprazolam e verapamil.

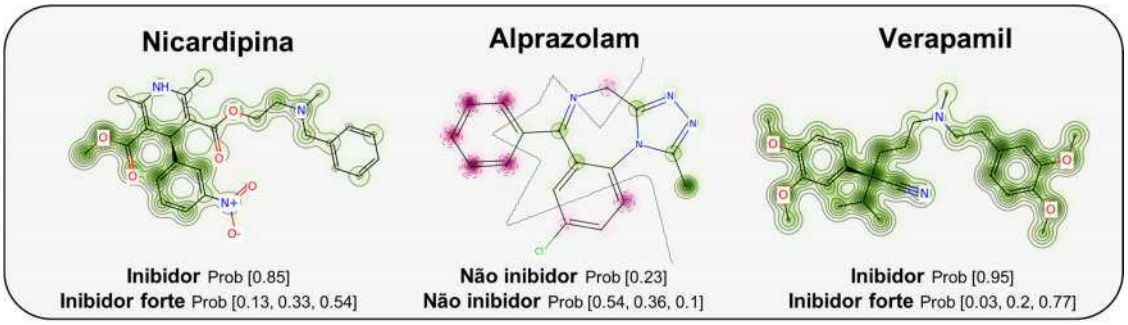

Figura 21. Mapa de probabilidade predita (MPP) para o modelo Morgan-RF de modelos binários de inibição de CYP34A. Bits verdes (átomos/fragmentos): contribuição favorável na propriedade (inibidor da CYP3A4); Bits rosa: contribuição desfavorável na propriedade (não inibidor de CYP3A4); Bits cinza: contribuição neutra na propriedade. Linhas em cinza delimitam a separação da contribuição desfavorável e favorável. O tamanho do vetor-bit de Morgan foi 1024 bits.

O anti-hipertensivo nicardipina é inibidor de CYP3A4 (MIYAJIMA et al., 2007). Foi classificado como inibidor pelo modelo binário, e pelo modelo multiclasse foi classificado como inibidor forte. As regiões delineadas com maior intensidade em sua estrutura são o anel fenila ligado ao grupo nitro, e a região do éter e cetona, esses fragmentos são considerados favoráveis para a propriedade.

O alprazolam é um fármaco benzodiazepínico utilizado como ansiolítico, hipnótico, pré-anestésico, anticonvulsivante e relaxante muscular, é substrato de CYP3A4 e não é inibidor da enzima (MOLANAEI et al., 2012). De acordo com a estrutura química do fármaco, pode-se observar que somente o anel triazol fusionado está delineado em verde com baixa intensidade, indicando que esse fragmento tem contribuição favorável para a propriedade. As linhas em cinzas demarcam a separação entre os fragmentos com características favoráveis e desaforáveis para a propriedade, como pode ser observado pelo anel aromático delineado em cor rosa, o que representa contribuição desfavorável para propriedade de inibição. O alprazolam foi classificado como não inibidor pelo modelo binário e multiclasse.

O verapamil é um fármaco antiarrítmico, utilizado em pacientes com arritmias cardíacas, é inibidor de CYP3A4 (DINGER; MEYER; MAURER, 2014). Foi classificado pelo modelo binário e multiclasse como inibidor de CYP3A4. Ao observar a estrutura do fármaco, pode-se observar que toda estrutura do fármaco, apresenta fragmentos delineados em cor verde, como nos dois anéis dimetoxi-metil-benzeno e a região da nitrila, indicando que todos esses fragmentos apresentam contribuição favorável para atividade de inibição de CYP3A4.

Com a geração dos MPPs pode-se observar a importância da interpretação mecanística dos modelos de QSAR, para encontrar relações entre os descritores e a atividade biológica ou propriedade, em via de se compreender melhor o mecanismo de ação de uma estrutura química ou aprofundar o conhecimento biológico sobre a propriedade em estudo.

5 CONCLUSÕES

- Foram integrados, preparados e balanceados os maiores conjuntos de dados publicamente disponíveis para substratos e inibidores de CYP3A4 com compostos químicos estruturalmente diversos.
- Os modelos de QSAR gerados para a identificação de substratos (modelos binários) e inibidores (modelos binários e multiclasse) apresentaram bons resultados estatísticos, tanto para validação externa e interna. Ao final destes resultados pode se comprovar a robustez e a alta capacidade preditiva dos modelos.
- Foram realizadas a análise de seis conjuntos de dados (A-F) com diferentes proporções de balanceamento. Dentre os seis tipos de balanceamentos, o conjunto B foi selecionado por apresentar os melhores resultados estatísticos e utilizado para a otimização dos modelos de QSAR para classificar corretamente os compostos em substratos e não substratos de CYP3A4.
- Os resultados dos parâmetros estatísticos dos modelos binários de substratos de CYP3A4 indicam que o modelo AtomPair-GBM apresentou maior sensibilidade e menor discrepância entre os valores de sensibilidade e especificidade. Além disso, o teste da randomização da variável Y demonstrou que as análises não são aleatórias, sugerindo que o modelo AtomPair-GBM é o mais adequado para classificar corretamente tanto os substratos quanto os não substratos de CYP3A4.
- Os resultados estatísticos para os modelos binários e multiclasse de inibidores de CYP3A4, apresentaram valores semelhantes quanto aos parâmetros estatísticos, sugerindo que o modelo Morgan-RF apresentou resultados ligeiramente superiores ao modelo Morgan-SVM.
- Com a geração dos MPPs pode ser observado à importância da interpretação mecanística dos modelos de QSAR, em via de se compreender a relação entre grupos e/ou fragmentos estruturais relevantes na inibição de CYP3A4. Estes modelos podem auxiliar na predição do perfil metabólico, como na interação fármaco-fármaco, de candidatos a fármacos nas etapas inicias de desenvolvimento de fármacos.
- Os modelos de substratos e inibidores de CYP3A4 estarão disponíveis para serem usados pela comunidade científica e indústrias farmacêuticas. Devido à sua robustez e consistência, estes modelos são guias úteis em Química Medicinal na

identificação, seleção e otimização de compostos candidatos a fármacos. Os modelos será disponibilizados no site do Laboratório de Planejamento de Fármacos e Modelagem Molecular - LabMol (http://labmol.farmacia.ufg.br).

6 REFERÊNCIAS BIBLIOGRÁFICAS

ADEUSI, S. **Pharmaceutical R & D : An Organizational Design Approach to Enhancing Productivity**. 2011. 112 f. Dissertação (Mestrado em Ciências em Gestão) -Instituto de tecnologia, Massachusetts.

ANDRADE, C. H.; SILVA, D. C.; BRAGA, R. C. In silico Prediction of Drug Metabolism by P450. **Current drug metabolism**, v. 15, n. 5, p. 514–525, 2014.

APPIAH-OPONG, R.; DE ESCH, I.; COMMANDEUR, J. N. M.; ANDARINI, M.; VERMEULEN, N. P. E. Structure-activity relationships for the inhibition of recombinant human cytochromes P450 by curcumin analogues. **European journal of medicinal chemistry**, v. 43, n. 8, p. 1621–1631, 2008.

ARIMOTO, R. Computational models for predicting interactions with cytochrome p450 enzyme. **Current topics in medicinal chemistry**, v. 6, n. 15, p. 1609–1618, 2006.

BAJORATH, J. Integration of virtual and high-throughput screening. **Nature reviews. Drug discovery**, v. 1, n. 11, p. 882–894, 2002.

BASKIN, I.; VARNEK, A. Building a chemical space based on fragment descriptors. **Combinatorial chemistry & high throughput screening**, v. 11, n. 8, p. 661–668, 2008.

BELFIELD, G. P.; DELANEY, S. J. The impact of molecular biology on drug discovery. **Biochemical Society transactions**, v. 34, p. 313–316, 2006.

BEN-DOR, A.; BRUHN, L.; FRIEDMAN, N.; NACHMAN, I.; SCHUMMER, M.; YAKHINI, Z. Tissue classification with gene expression profiles. **Journal of computational biology : a journal of computational molecular cell biology**, v. 7, n. 3-4, p. 559–583, 2000.

BERNHARDT, R. Cytochromes P450 as versatile biocatalysts. **Journal of biotechnology**, v. 124, n. 1, p. 128–145, 2006.

BARREIRO, E.J.; FRAGA, C. A. M. **Química medicinal: as bases moleculares da ação dos fármacos**, 2ª Ed. Porto Alegre, editora Artmed. 2008.

BRAGA, R. C. ; ALVES, V. M. ; SILVA, F. C. ; ANDRADE, C. H. **QSAR and Molecular Modeling Approaches for Prediction of Drug Metabolism.** In: Alexander Lyubimov. (Org.). Encyclopedia of drug metabolism and interactions. Part VI. Methods and Protocols for Prediction and Evaluation of Drug Metabolism and Drug Interaction Studies. 1ed. Hoboken, NJ, USA: John Wiley & Sons, Inc, 2015, v. 7, p. 1-28.

BRAGA, R. C.; ANDRADE, C. H. QSAR and QM/MM approaches applied to drug metabolism prediction. **Mini reviews in medicinal chemistry**, v. 12, n. 6, p. 573–582, 2012.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. Belmont: Wadsworth Publishing, 1984.

BUCHWALD, P.; YAMASHITA, F. Bilinear Model for the Size-Dependency of the CYP3A4 Inhibitory Activity of Structurally Diverse Compounds. **Molecular Informatics**, v. 33, p. 8–14, 2014.

BUCKLE, D. R.; ERHARDT, P. W.; GANELLIN, C. R.; KOBAYASHI, T.; PERUN, T. J.; PROUDFOOT, J.; SENN-BILFINGER, J. Glossary of terms used in medicinal chemistry Part II (IUPAC recommendations 2013). **Annual reports in medicinal chemistry**, v. 48, n. 8, p. 386–418, 2013.

BURTON, J.; PETIT, J.; DANLOY, E.; MAGGIORA, G. M.; VERCAUTEREN, D. P. Rough Set Theory as an Interpretable Method for Predicting the Inhibition of Cytochrome P450 1A2 and 2D6. **Molecular Informatics**, v. 32, n. 7, p. 579–589, 2013.

CARBON-MANGELS, M.; HUTTER, M. C. Selecting Relevant Descriptors for Classification by Bayesian Estimates: A Comparison with Decision Trees and Support Vector Machines Approaches for Disparate Data Sets. **Molecular Informatics**, v. 30, n. 10, p. 885–895, 2011.

CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. Atom Pairs as Molecular Features in Structure-Activity Studies : Definition and Applications. **American chemical society**, v. 13, n. 4, p. 8–11, 1985.

CHERKASOV, A.; MURATOV, E. N.; FOURCHES, D.; VARNEK, A.; BASKIN, I. I.; CRONIN, M.; DEARDEN, J.; GRAMATICA, P.; MARTIN, Y. C.; TODESCHINI, R.; CONSONNI, V.; KUZ'MIN, V. E.; CRAMER, R.; BENIGNI, R.; YANG, C.; RATHMAN, J.; TERFLOTH, L.; GASTEIGER, J.; RICHARD, A.; TROPSHA, A. QSAR Modeling: Where Have You Been? Where Are You Going To? **Journal of medicinal chemistry**, v. 57, n. 12, p. 4977–5010, 2014.

COHEN, F. J. Macro trends in pharmaceutical innovation. **Nature reviews. drug discovery**, v. 4, n. 1, p. 78–84, 2005.

COHEN, N. C.; KOEHLER, K. F.; RAO, S. N.; SNYDER, J. P.; COHEN, N. C.; ITAI, A.; MIZUTANI, M. Y.; NISHIBATA, Y.; TOMIOKA, N.; TOLLENAERE, J. P.; GUND, P.; MAGGIORA, G.; SNYDER, J. P.; PRIESTLE, J. P.; PARIS, C. G.; COHEN, N. C.; HUBBARD, R. E.; GUND, T. **Guidebook on Molecular Modeling in Drug Design**. San Diego: Academic Press, 1996. v. 361

CONSONNI, V.; BALLABIO, D.; TODESCHINI, R. Comments on the definition of the Q2 parameter for QSAR validation. **Journal of chemical information and modeling**, v. 49, n. 7, p. 1669–1678, 2009.

CONSONNI, V.; TODESCHINI, R. Molecular descriptors. In: PUZYN, T.; LESZCZYNSKI, J.; CRONIN, M. T. (Eds). **Recent Advances in QSAR Studies**. Dordrecht: Springer, 2010.

CRAMER, R. D. The inevitable QSAR renaissance. **Journal of computer-aided molecular design**, v. 26, n. 1, p. 35–38, 2012.

CRAMER, R. D.; PATTERSON, D. E.; BUNCE, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. **Journal of the American Chemical Society**, v. 110, n. 18, p. 5959–5967, 1988.

CRONIN, M. T. D. Quantitative structure-activity relationships (QSARs) -- aplications and methodology. In: PUZYN, T.; LESZCZYNSKI, J.; CRONIN, M. T. (Eds.). . **Recent Advances in QSAR Studies**. Dordrecht: Springer, 2010. p. 3–11.

CROOKE, S. T. Optimizing the impact of genomics on drug discovery and development. **Nature biotechnology**, v. 16 Suppl, p. 29–30, 1998.

DALKAS, G. A; VLACHAKIS, D.; TSAGKRASOULIS, D.; KASTANIA, A.; KOSSIDA, S. State-of-the-art technology in modern computer-aided drug design. **Briefings in bioinformatics**, v. 14, n. 6, p. 745–752, 2013.

DEBNATH, A. K. Quantitative structure-activity relationship (QSAR) paradigm--Hansch era to new millennium. **Mini reviews in medicinal chemistry**, v. 1, n. 2, p. 187–195, 2001.

DIDZIAPETRIS, R.; DAPKUNAS, J.; SAZONOVAS, A.; JAPERTAS, P. Trainable structure-activity relationship model for virtual screening of CYP3A4 inhibition. **Journal of computer-aided molecular design**, v. 24, n. 11, p. 891–906, 2010.

DIMASI, J. A; HANSEN, R. W.; GRABOWSKI, H. G. The price of innovation: new estimates of drug development costs. **Journal of health economics**, v. 22, n. 2, p. 151–185, 2003.

DINGER, J.; MEYER, M. R.; MAURER, H. H. Development of an in vitro cytochrome P450 cocktail inhibition assay for assessing the inhibition risk of drugs of abuse. **Toxicology letters**, v. 230, n. 1, p. 28–35, 2014.

DOSHI, U.; LI, A. P. Luciferin IPA-Based Higher Throughput Human Hepatocyte Screening Assays for CYP3A4 Inhibition and Induction. **Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening**, v. 16, p. 903–909, 2011.

DUAN, J.; DIXON, S. L.; LOWRIE, J. F.; SHERMAN, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. **Journal of molecular graphics & modelling**, v. 29, n. 2, p. 157–170, 2010.

EITRICH, T.; KLESS, A.; DRUSKA, C.; MEYER, W.; GROTENDORST, J. Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques. **Journal of chemical information and modeling** , v 47, n. 92, p. 92–103, 2007.

EKINS, S.; HONEYCUTT, J. D.; METZ, J. T. Evolving molecules using multi-objective optimization: applying to ADME/Tox. **Drug discovery today**, v. 15, n. 11-12, p. 451–460, 2010.

EWING, T.; FEHER, M. Forecasting CYP2D6 and CYP3A4 Risk with a Global/Local Fusion Model of CYP450 Inhibition. **Molecular Informatics**, v. 29, n. 1-2, p. 127–141, 2010.

FDA. Guidance for Industry, Investigators, and Reviewers Exploratory IND Studies. **FDA Food and Drug Administration**, p. 13 p, 2006.

FERNÁNDEZ, A. G.; GARCÍA, M. A. DE S.; FERNÁNDEZ, A. M. M.; RAMOS, S. B.; GALÁN, M. J. G. **Aspectos fundamentales del Citocromo P450** (F. B. Moya, Ed.)Madrid.ADEMAS Comunicación Gráfica, S.l., , 2009.

FOTI, R. S.; ROCK, D. A; HAN, X.; FLOWERS, R. A; WIENKERS, L. C.; WAHLSTROM, J. L. Ligand-based design of a potent and selective inhibitor of cytochrome P450 2C19. **Journal of Medicinal Chemistry**, v. 55, n. 3, p. 1205–1214, 2012.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. **Journal of chemical information and modeling**, v. 50, n. 7, p. 1189–1204, 2010.

GRAMATICA, P. Principles of QSAR models validation: internal and external. **QSAR & Combinatorial Science**, v. 26, n. 5, p. 694–701, 2007.

GUENGERICH, F. P. Cytochrome P450s and other enzymes in drug metabolism and toxicity. **The AAPS journal**, v. 8, n. 1, p. E101–E111, 2006.

GUIDO, R. V. C.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. **Estudos Avançados**, v. 24, n. 70, p. 81–98, 2010.

GUIDO, R. V. C.; OLIVA, G.; ANDRICOPULO, A. D. Modern drug discovery technologies: opportunities and challenges in lead discovery. **Combinatorial chemistry & high throughput screening**, v. 14, n. 10, p. 830–839, 2011.

HAJI-MOMENIAN, S.; RIEGER, J. M.; MACDONALD, T. L.; BROWN, M. L. Comparative molecular field analysis and QSAR on substrates binding to cytochrome p450 2D6. **Bioorganic & medicinal chemistry**, v. 11, n. 24, p. 5545–5554, 2003.

HANDA, K.; NAKAGOME, I.; YAMAOTSU, N.; GOUDA, H.; HIRONO, S. Three-dimensional Quantitative Structure-Activity Relationship Analysis of Inhibitors of Human and Rat Cytochrome P4503A Enzymes. **Drug Metabolism and Pharmacokinetics**, v. 28, n. 4, p. 345–355, 2013.

HANSCH, C. Quantitative approach to biochemical structure-activity relationships. **Accounts of Chemical Research**, v. 2, n. 8, p. 232–239, 1969.

HANSCH, C.; FUJITA, T. p -σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. **Journal of the American Chemical Society**, v. 86, n. 8, p. 1616–1626, 1964.

HO, H. K.; CHAN, J. C. Y.; HARDY, K. D.; CHAN, E. C. Y. Mechanism-based inactivation of CYP450 enzymes: a case study of lapatinib. **Drug Metabolism Reviews**, n. 2, p. 1–8, 2015.

HOPFINGER, A. J.; WANG, S.; TOKARSKI, J. S.; JIN, B. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. **Journal Americam Chemical Society**, v. 7863, n. 5, p. 10509–10524, 1997.

HUGHES, J. P.; REES, S.; KALINDJIAN, S. B.; PHILPOTT, K. L. Principles of early drug discovery. **British Journal of Pharmacology**, v. 162, n. 6, p. 1239–1249, 2011.

ITOKAWA, D.; NISHIOKA, T.; FUKUSHIMA, J.; YASUDA, T.; YAMAUCHI, A.; CHUMAN, H. Quantitative Structure–Activity Relationship Study of Binding Affinity of Azole Compounds with CYP2B and CYP3A. **QSAR & Combinatorial Science**, v. 26, n. 7, p. 828–836, 2006.

ITOKAWA, D.; YAMAUCHI, A.; CHUMAN, H. Quantitative Structure-Activity Relationship for Inhibition of CYP2B6 and CYP3A4 by Azole Compounds - Comparison with Their Binding Affinity. **QSAR & Combinatorial Science**, v. 28, n. 6-7, p. 629–636, 2007.

JENSEN, B. F.; VIND, C.; PADKJÆR, S. B.; BROCKHOFF, P. B.; REFSGAARD, H. H. F. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k -Nearest Neighbor and Extended Connectivity Fingerprints , Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. **Journal medicinal chemistry**, v. 50, p. 501–511, 2007.

JOHN M. BEALE, J.; BLOCK, J. H. **Organic Medicinal and Pharmaceutical Chemistry**. 12. ed. Philadelphia, PA: Lippincott Williams & Wilkinks, 2011.

JOHNSTON, W. A; HUANG, W.; DE VOSS, J. J.; HAYES, M. A; GILLAM, E. M. J. Quantitative whole-cell cytochrome P450 measurement suitable for high-throughput application. **Journal of biomolecular screening**, v. 13, n. 2, p. 135–141, 2008.

JÓNSDÓTTIR, S. Ó.; RINGSTED, T.; NIKOLOV, N. G.; DYBDAHL, M.; WEDEBYE, E. B.; NIEMELÄ, J. R. Identification of cytochrome P450 2D6 and 2C9 substrates and inhibitors by QSAR analysis. **Bioorganic & medicinal chemistry**, v. 20, n. 6, p. 2042–2053, 2012.

JULIANO, R. L. Pharmaceutical innovation and public policy: The case for a new strategy for drug discovery and development. **Science and public policy**, v. 40, n. 3, p. 393–405, 2013.

KHOJASTEH, S. C.; PRABHU, S.; KENNY, J. R.; HALLADAY, J. S.; LU, A. Y. H. Chemical inhibitors of cytochrome P450 isoforms in human liver microsomes: a re-evaluation of P450 isoform selectivity. **European journal of drug metabolism and pharmacokinetics**, v. 36, n. 1, p. 1–16, 2011.

KIRCHMAIR, J.; WILLIAMSON, M. J.; TYZACK, J. D.; TAN, L.; BOND, P. J.; BENDER, A.; GLEN, R. C. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. **Journal of chemical information and modeling**, v. 52, n. 3, p. 617–648, 2012.

KORHONEN, L. E.; TURPEINEN, M.; RAHNASTO, M.; WITTEKINDT, C.; POSO, A.; PELKONEN, O.; RAUNIO, H.; JUVONEN, R. O. New potent and selective cytochrome P450 2B6 (CYP2B6) inhibitors based on three-dimensional quantitative structure-activity relationship (3D-QSAR) analysis. **British journal of pharmacology**, v. 150, n. January, p. 932–942, 2007.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. **Artificial intelligence review**, v. 26, n. 3, p. 159–190, 2007.

KUBINYI, H. QSAR and 3D QSAR in drug design Part 1: methodology. **Drug discovery today**, v. 2, n. 11, p. 457–467, 1997.

KUMAR, G. N.; SURAPANENI, S. Role of drug metabolism in drug discovery and development. **Medicinal research reviews**, v. 21, n. 5, p. 397–411, 2001.

LEWIS, D. F. V; LAKE, B. G.; DICKINS, M. Quantitative structure-activity relationships (QSars) in CYP3A4 inhibitors: the importance of lipophilic character and hydrogen bonding. **Journal of Enzyme Inhibition and Medicinal Chemistry**, v. 21, n. 2, p. 127–132, 2006.

LI, D. Q.; KIM, R.; MCARTHUR, E.; FLEET, J. L.; BAILEY, D. G.; JUURLINK, D.; SHARIFF, S. Z.; GOMES, T.; MAMDANI, M.; GANDHI, S.; DIXON, S.; GARG, A. X. Risk of adverse events among older adults following co-prescription of

clarithromycin and statins not metabolized by cytochrome P450 3A4. **Canadian Medical Association Journal**, v. 187, n. 3, p. 174–180, 2014.

LI, H.; SUN, J.; FAN, X.; SUI, X.; ZHANG, L.; WANG, Y.; HE, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. **Journal of computer-aided molecular design**, v. 22, n. 11, p. 843–855, 2008.

LIMA, L. M. Fundamentos do Metabolismo de Fármacos. In: J.BARREIRO, E.; FRAGA, C. A. M. (Eds.). . **Química Medicinal: As bases moleculares da ação dos fármacos**. 3. ed. Porto Alegre: Artmed S.A, 2015. p. 43–103.

LOMBARDINO, J. G.; LOWE, J. A. The role of the medicinal chemist in drug discovery--then and now. **Nature reviews. Drug discovery**, v. 3, n. 10, p. 853–862, 2004.

MANUSCRIPT, A.; CLIFFORD, D. B.; EVANS, S.; YANG, Y.; ACOSTA, E. P.; RIBAUDO, H.; GULICK, R. M.; TEAM, A. S.; HISTORY, D.; ORIGINS, M. H.; MAPPING, D.; ORRELL, C.; COHEN, K.; CONRADIE, F.; ZEINECKER, J.; IVE, P.; SANNE, I.; WOOD, R.; LAKHMAN, S. S.; MA, Q.; MORSE, G. D. Pharmacogenomics of CYP3A: considerations for HIV treatment. **Pharmacogenomics**, v. 6, n. 8, p. 343–355, 2010.

MARTINEZ-SANZ, J.; BONNET, P.; LOZANO, S.; ARRAULT, A.; MORIN-ALLORY, L.; VAYER, P. New QSAR Models for Human Cytochromes P450, 1A2, 2D6 and 3A4 Implicated in the Metabolism of Drugs. Relevance of Dataset on Model Development. **Molecular Informatics**, v. 32, n. 7, p. 573–577, 2013.

MAYHEW, B. S.; JONES, D. R.; HALL, S. D. AN IN VITRO MODEL FOR PREDICTING IN VIVO INHIBITION OF CYTOCHROME P450 3A4 BY METABOLIC INTERMEDIATE COMPLEX FORMATION ABSTRACT : **Drug metabolism and disposition**, v. 28, n. 9, p. 1031–1037, 2000.

MCGEE, P. Clinical trial on the move. **Drug Discovery & Development**, v. 9, n. 6, p. 16–22, 2006.

MELVILLE, J. L.; BURKE, E. K.; HIRST, J. D. Machine learning in virtual screening. **Combinatorial chemistry & high throughput screening**, v. 12, n. 4, p. 332–343, 2009.

MISHRA, N. K.; AGARWAL, S.; RAGHAVA, G. P. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. **BMC pharmacology**, v. 10, p. 8, 2010.

MISHRA, P.; TRIPATHI, V.; YADAV, B. S. Insilco QSAR modeling and drug development process. v. 1, n. December, p. 37–40, 2010.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill Science, 1997.

MIYAJIMA, S.; NEMOTO, K.; SEKIMOTO, M.; KINAE, Y.; KASAHARA, T.; SOUMA, S.; DEGAWA, M. Induction of hepatic cytochrome P450 isoforms by nicardipine at therapeutic doses in spontaneously hypertensive rats. **The Journal of toxicological sciences**, v. 32, n. 1, p. 79–90, 2007.

MODA, T. L. **Desenvolvimento de Modelos In Silico de Propriedades de ADME Para a Triagem de Novos Candidatos a Fármacos.** 2007. 97 f. Dissertação (Mestrado em Ciências: Física e Aplicada- Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos.

MODA, T. L.; TORRES, L. G.; CARRARA, A. E.; ANDRICOPULO, A. D. PK/DB: database for pharmacokinetic properties and predictive in silico ADME models. **Bioinformatics**, v. 24, n. 19, p. 2270–2271, 2007.

MOLANAEI, H.; STENVINKEL, P.; QURESHI, A. R.; CARRERO, J. J.; HEIMBÜRGER, O.; LINDHOLM, B.; DICZFALUSY, U.; ODAR-CEDERLÖF, I.; BERTILSSON, L. Metabolism of alprazolam (a marker of CYP3A4) in hemodialysis patients with persistent inflammation. **European Journal of Clinical Pharmacology**, v. 68, n. 5, p. 571–577, 2012.

MONTELLANO, P. R. O. DE. **Cytochrome P450: Structure, Mechanism, and Biochemistry**. New York: Springer, 2010.

MORGAN, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. **Journal of Chemical Documentation**, v. 5, n. 2, p. 107–113, 1965.

MULLARD, A. New drugs cost US$2.6 billion to develop. **Nature reviews. Drug discovery**, v. 13, n. 12, p. 877, 2014.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, v. 7, n. December, p. 21, 2013.

NATURE. End of the Lipitor Era. **Nature Reviews Drug Discovery**, v. 10, n. 12, p. 889–889, 2011.

NELSON, D. R. Progress in tracing the evolutionary paths of cytochrome P450. **Biochimica et biophysica acta**, v. 1814, n. 1, p. 14–18, 2011.

NICOLAOU, K. C. Advancing the Drug Discovery and Development Process. **Angewandte Chemie International Edition**, p. n/a – n/a, 2014.

OECD. OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models. . 2004, p. 1–2.

OMURA, T.; SATO, R. The carbon monoxide-bindng pigent of liver microsomes. I. Evidence for its hemoprotein nature. **The Journal of biological chemistry**, v. 239, n. 7, p. 2370–2378, 1964.

OOMS, F. Molecular modeling and computer aided drug design . Examples of their applications in medicinal chemistry. **Current medicinal chemistry**, p. 141–158, 2000.

OPREA, T. I.; MATTER, H. Integrating virtual screening in lead discovery. **Current opinion in chemical biology**, v. 8, n. 4, p. 349–358, 2004.

OWENS, J. Enigmatic enzyme. **Nature Reviews Drug Discovery**, v. 5, n. 11, p. 893–893, 2006.

PAMMOLLI, F.; MAGAZZINI, L.; RICCABONI, M. The productivity crisis in pharmaceutical R&D. **Nature reviews. Drug discovery**, v. 10, n. 6, p. 428–438, 2011.

PELKONEN, O.; TURPEINEN, M.; HAKKOLA, J.; HONKAKOSKI, P.; HUKKANEN, J.; RAUNIO, H. Inhibition and induction of human cytochrome P450 enzymes: current status. **Archives of toxicology**, v. 82, n. 10, p. 667–715, 2008.

PHRMA. **Drug Discovery and Development: Overview.**Washington, DC.John Wiley & Sons, Inc, , 2007.

PHRMA. **Pharmaceutical Research and Manufacturers of America.**Washington, DC, 2013.

PIRMOHAMED, M.; PARK, B. K. Cytochrome P450 enzyme polymorphisms and adverse drug reactions. **Toxicology**, v. 192, n. 1, p. 23–32, 2003.

POWERS, D. M. . Evaluation: from precision, recall and F-measure to roc informedness, markedness & correlation. **Jornal of Machine Learning Tecnologies**, v. 2, n. 1, p. 37–63, 2011.

RAHNASTO, M.; RAUNIO, H.; POSO, A.; WITTEKINDT, C.; JUVONEN, R. O. Quantitative structure-activity relationship analysis of inhibitors of the nicotine metabolizing CYP2A6 enzyme. **Journal of medicinal chemistry**, v. 48, n. 2, p. 440–449, 2005.

RANDIC, M.; BASAK, S. C. A New Descriptor for Structure-Property and Structure-Activity Correlations. **Journal of Chemical Information and Modeling**, v. 41, n. 3, p. 650–656, 2001.

RESFSGAARD, H. H. F.; JENSEN, B. F.; CHRISTENSEN, I. T.; HAGEN, N.; BROCKHOFF, P. B. In Silico Prediction of Cytochrome P450 Inhibitors. **Drug Development Research**, v. 429, n. April, p. 417–429, 2006.

ROY, K.; PRATIM ROY, P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. **European journal of medicinal chemistry**, v. 44, n. 7, p. 2913–2922, 2009.

ROY, K.; ROY, P. P. Exploring QSAR and QAAR for inhibitors of cytochrome P450 2A6 and 2A5 enzymes using GFA and G / PLS techniques. **European Journal of Medicinal Chemistry**, v. 44, n. 5, p. 1941–1951, 2008.

RÜCKER, C.; RÜCKER, G.; MERINGER, M. y-Randomization and its variants in QSPR/QSAR. **Journal of chemical information and modeling**, v. 47, n. 6, p. 2345–2357, 2007.

SARACENO, M.; MASSARELLI, I.; IMBRIANI, M.; JAMES, T. L.; BIANUCCI, A. M. Optimizing QSAR models for predicting ligand binding to the drug-metabolizing cytochrome P450 isoenzyme CYP2D6. **Chemical biology & drug design**, v. 78, n. 2, p. 236–251, 2011.

SEOANE, J. A; LÓPEZ-CAMPOS, G.; DORADO, J.; MARTIN-SANCHEZ, F. New approaches in data integration for systems chemical biology. **Current topics in medicinal chemistry**, v. 13, n. 5, p. 591–601, 2013.

SHITYAKOV, S.; PUSKÁS, I.; ROEWER, N.; FÖRSTER, C.; BROSCHEIT, J. Three-dimensional quantitative structure-activity relationship and docking studies in a series of anthocyanin derivatives as cytochrome P450 3A4 inhibitors. **Advances and applications in bioinformatics and chemistry : AABC**, v. 7, p. 11–21, 2014.

SMITH, D. A; DALVIE, D. Why do metabolites circulate? **Xenobiotica; the fate of foreign compounds in biological systems**, v. 44, n. September, p. 1–20, 2011.

SNYDER, R.; SANGAR, R.; WANG, J.; EKINS, S. Three-Dimensional Quantitative Structure Activity Relationship for Cyp2d6 Substrates. **Molecular Informatics**, v. 21, n. 4, p. 357–368, 2002.

STEINBECK, C.; HAN, Y.; KUHN, S.; HORLACHER, O.; LUTTMANN, E.; WILLIGHAGEN, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. **Journal of chemical information and computer sciences**, v. 43, n. 2, p. 493–500, 2003.

STJERNSCHANTZ, E.; VERMEULEN, N. P. E.; OOSTENBRINK, C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. **Expert opinion on drug metabolism & toxicology**, v. 4, n. 5, p. 513–527, 2008.

STROHMEIER, G. A.; PICHLER, H.; MAY, O.; GRUBER-KHADJAWI, M. Application of designed enzymes in organic synthesis. **Chemical reviews**, v. 111, n. 7, p. 4141–4164, 2011.

SUKUMAR, N.; DAS, S. Current trends in virtual high throughput screening using ligand-based and structure-based methods. **Combinatorial chemistry & high throughput screening**, v. 14, n. 10, p. 872–888, 2011.

SUN, H.; VEITH, H.; XIA, M.; AUSTIN, C. P.; TICE, R. R.; HUANG, R. Prediction of Cytochrome P450 Profiles of Environmental Chemicals with QSAR Models Built from Drug-like Molecules. **Molecular informatics**, v. 31, n. 11-12, p. 783–792, 2012.

SUTTON, D.; BUTLER, A. M.; NADIN, L.; MURRAY, M. Role of CYP3A4 in human hepatic diltiazem N-demethylation: inhibition of CYP3A4 activity by oxidized diltiazem metabolites. **The Journal of pharmacology and experimental therapeutics**, v. 282, n. 1, p. 294–300, 1997.

TAAVITSAINEN, P. **Cytochrome P450 isoform-specific in vitro methods to predict drug metabolism and interactions**. [s.l.] University of Oulu, 2001.

TAO ZHANG, QI CHEN, LI LI, LIMIM ANGELA, L. D.-Q. W. In silico methods for prediction of drug metabolism. **Combinatorial chemistry & high throughput screening**, v. 14, p. 388–395, 2011.

TAVARES, L. C. QSAR: A ABORDAGEM DE HANSCH. **Química Nova**, v. 27, n. 4, p. 631–639, 2004.

TERFLOTH, L.; BIENFAIT, B.; GASTEIGER, J. Ligand-Based Models for the Isoform Specificity of Cytochrome P450 3A4, 2D6, and 2C9 Substrates. **Journal of chemical information and modeling**, n. 47, p. 1688–1701, 2007.

TESTA, B.; PEDRETTI, A.; VISTOLI, G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. **Drug Discovery Today**, v. 17, n. 11-12, p. 549–560, 2012.

THOMSOM REUTERES. **2012 CMR INTERNATIONAL PHARMACEUTICAL R & D FACTBOOK**, 2012.

TODESCHINI, R.; CONSONNI, V. **Handbook of molecular descriptors**. Weinheim, Germany: Wiley-VCH Verlag GmbH, 2008.

TODESCHINI, R.; CONSONNI, V. **Molecular Descriptors for Chemoinformatics**. Weinheim, Germany: Wiley-VCH, 2009.

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

TROPSHA, A.; GOLBRAIKH, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. **Current pharmaceutical design**, v. 13, n. 34, p. 3494–3504, 2007.

UENG, Y.-F.; CHEN, C.-C.; YAMAZAKI, H.; KIYOTANI, K.; CHANG, Y.-P.; LO, W.-S.; LI, D.-T.; TSAI, P.-L. Mechanism-based Inhibition of CYP1A1 and CYP3A4 by the Furanocoumarin Chalepensin. **Drug Metabolism and Pharmacokinetics**, v. 28, n. 3, p. 229–238, 2013.

VAN DE WATERBEEMD, H.; GIFFORD, E. ADMET in silico modelling: towards prediction paradise? **Nature reviews. Drug discovery**, v. 2, n. 3, p. 192–204, 2003.

VARNEK, A.; BASKIN, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. **Molecular Informatics**, v. 30, n. 1, p. 20–32, 2011.

VAZ, R. J.; NAYEEM, A.; SANTONE, K.; CHANDRASENA, G.; GAVAI, A. V. A 3D-QSAR model for CYP2D6 inhibition in the aryloxypropanolamine series. **Bioorganic & medicinal chemistry letters**, v. 15, n. 17, p. 3816–3820, 2005.

VEDANI, A.; DOBLER, M. 5D-QSAR: the key for simulating induced fit? **Journal of medicinal chemistry**, v. 45, n. 11, p. 2139–2149, 2002.

VEDANI, A.; DOBLER, M.; LILL, M. A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. **Journal of medicinal chemistry**, v. 48, n. 11, p. 3700–3703, 2005.

VEERASAMY, R.; RAJAK, H.; JAIN, A.; SIVADASAN, S.; VARGHESE, C. P.; AGRAWAL, R. K. Validation of QSAR Models - Strategies and Importance. **International Journal and Drug Design and Discovery**, v. 2, n. 3, p. 511–519, 2011.

VERLI, H.; ALBUQUERQUE, M. G.; BICCA DE ALENCASTRO, R.; BARREIRO, E. J. Local intersection volume: a new 3D descriptor applied to develop a 3D-QSAR pharmacophore model for benzodiazepine receptor ligands. **European Journal of Medicinal Chemistry**, v. 37, n. 3, p. 219–229, 2002.

VILLAGRA, D.; GOETHE, J.; SCHWARTZ, H. I.; SZAREK, B.; KOCHERLA, M.; GOROWSKI, K.; WINDEMUTH, A.; RUAÑO, G. Novel drug metabolism indices for pharmacogenetic functional status based on combinatory genotyping of CYP2C9, CYP2C19 and CYP2D6 genes. **Biomarkers in Medicine**, v. 5, n. 4, p. 427–438, 2011.

WALTERS, W. P.; GREEN, J.; WEISS, J. R.; MURCKO, M. A. What do medicinal chemists actually make? A 50-year retrospective. **Journal of Medicinal Chemistry**, v. 54, n. 19, p. 6405–6416, 2011.

WELLING, M. A First Encounter with Machine Learning. **Donald Bren School of Information and Computer Science**. Irvine, CA: University of California, 2010.

WERMUTH, C.; GANELLIN, C.; LINDBERG, P. GLOSSARY OF TERMS USED IN MEDICINAL CHEMISTRY (IUPAC Recommendations 1998). **Pure and Applied Chemistry**, v. 70, n. 5, p. 1129–1143, 1998.

WILKINS, M. R.; PASQUALI, C.; APPEL, R. D.; OU, K.; GOLAZ, O.; SANCHEZ, J. C.; YAN, J. X.; GOOLEY, A. A.; HUGHES, G.; HUMPHERY-SMITH, I.; WILLIAMS, K. L.; HOCHSTRASSER, D. F. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. **Bio/technology (Nature Publishing Company)**, v. 14, n. 1, p. 61–65, 1996.

WILSON, G. L.; LILL, M. A. Integrating structure-based and ligand-based approaches for computational drug design. **Future Medicinal Chemistry**, v. 3, n. 6, p. 735–750, 2011.

WISHART, D. S. Applications of metabolomics in drug discovery and development. **Drugs in R&D**, v. 9, n. 5, p. 307–322, 2008.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining: Pratical Machine Learning Tools and Techniques**. 3. ed. San Francisco: Morgan Kaufmann, 2011.

XU, J.; HAGLER, A. Chemoinformatics and Drug Discovery. **Molecules**, v. 7, n. 8, p. 566–600, 2002.

YANG, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. **Drug Discovery Today**, v. 15, n. 11-12, p. 444–450, 2010.

YAP, C. W.; CHEN, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. **Journal of chemical information and modeling**, v. 45, n. 4, p. 982–992, 2005.

YASUO, K.; YAMAOTSU, N.; GOUDA, H.; TSUJISHITA, H.; HIRONO, S. Structure-based CoMFA as a predictive model - CYP2C9 inhibitors as a test case. **Journal of Chemical Information and Modeling**, v. 49, n. 4, p. 853–864, 2009.

ZANETTI, C. A. T. **Aplicação de Espectroscopia no Infravermelho Médio e Análise Discriminante por Mínimos Quadrados Parciais para Classificação de Biodieseis e Misturas Biodiesel/Diesel**. 2014. 105 f. Dissertação (Mestrado em Química) - Instituto de Química, Universidade Federal de Uberlândia, Uberlândia.

ZARETZKI, J.; BERGERON, C.; HUANG, T.; RYDBERG, P.; SWAMIDASS, S. J.; BRENEMAN, C. M. RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on drug-like molecules. **Bioinformatics (Oxford, England)**, v. 29, n. 4, p. 497–498, 2013.

ZARETZKI, J.; RYDBERG, P.; BERGERON, C.; BENNETT, K. P.; OLSEN, L.; BRENEMAN, C. M. RS-Predictor models augmented with SMARTCyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. **Journal of chemical information and modeling**, v. 52, n. 6, p. 1637–1659, 2012.

7 APÊNDICE I

7.1.1 Conjunto A

Tabela 5. Características estatísticas de modelos de QSAR para CYP3A4 avaliado por *5-fold* para o conjunto modelagem do conjunto A.

	Conjunto modelagem (6.570*)*										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACSS-PLSDA	0,06	0,00	0,00	0,00	0,06	0,00	1,00	0,11	0,50	0,69	ncomp=25
MACCS-SVM[1]	0,95	0,83	0,34	0,40	0,71	0,95	0,24	0,36	0,62	0,69	C=0,25
MACSS-SVM[2]	0,94	0,67	0,01	0,05	0,40	0,94	0,01	0,01	0,50	0,69	C=8192
MACCS-KNN	0,95	0,78	0,34	0,37	0,61	0,96	0,26	0,36	0,62	0,69	K=5
MACCS- CTREE[1]	0,94	0,73	0,21	0,26	0,51	0,95	0,15	0,23	0,57	0,69	mincri=0,378
MACCS-CTREE[2]	0,94	0,68	0,20	0,23	0,41	0,95	0,16	0,23	0,57	0,69	maxd=8
MACSS-GBM	0,95	0,79	0,45	0,46	0,61	0,96	0,39	0,47	0,69	0,69	interaction depth = 18, ntrees=1000, shrinkage=0,1
MACSS-RPART	0,94	0,69	0,23	0,26	0,44	0,95	0,18	0,26	0,58	0,69	cp=0,001

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM[1]: svm radialCost; SVM[2]: svm linear; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo; CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

Tabela 6. Características estatísticas de modelos de QSAR para CYP3A4 atribuído para o conjunto teste do Conjunto A.

	Conjunto teste (1.642)									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-PLSDA	0,05	0,00	0,00	0,00	0,05	0,00	1,00	0,10	0,50	0,60
MACSS-SVM[1]	0,96	0,82	0,36	0,40	0,68	0,96	0,26	0,38	0,63	0,60
MACSS-SVM[2]	0,05	0,00	0,00	0,00	0,05	0,00	1,00	0,10	0,50	0,60
MACCS-KNN	0,95	0,72	0,33	0,34	0,48	0,96	0,28	0,35	0,63	0,60
MACSS-CTREE[1]	0,95	0,71	0,25	0,28	0,45	0,96	0,20	0,28	0,59	0,60
MACSS, CTREE[2]	0,95	0,72	0,28	0,30	0,48	0,96	0,22	0,30	0,60	0,60
MACSS-GBM	0,95	0,76	0,43	0,44	0,56	0,97	0,38	0,45	0,68	0,60
MACCS-RPART	0,94	0,69	0,25	0,26	0,42	0,96	0,20	0,27	0,59	0,60

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM[1]: svm radialCost; SVM[2]: svm linear; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo; CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS Keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

7.1.2 Conjunto B

Tabela 7. Características estatísticas de modelos de QSAR para CYP3A4 avaliado por *5-fold* para o conjunto modelagem do conjunto B.

	Conjunto modelagem (760*)*										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACSS-PLSDA	0,76	0,75	0,39	0,42	0,74	0,76	0,43	0,54	0,68	0,73	ncomp=3
MACCS-SVM	0,80	0,78	0,53	0,54	0,74	0,83	0,63	0,68	0,76	0,73	C=4
MACCS-KNN	0,76	0,77	0,40	0,44	0,77	0,76	0,42	0,54	0,68	0,73	K=15
MACCS- CTREE[1]	0,75	0,73	0,38	0,40	0,70	0,76	0,44	0,54	0,67	0,73	mincri=0,949
MACCS-CTREE[2]	0,75	0,71	0,41	0,41	0,63	0,79	0,57	0,60	0,70	0,73	maxd=5
MACSS-GBM	0,80	0,78	0,55	0,55	0,71	0,84	0,68	0,70	0,77	0,73	shrinkage=0,1
MACSS-RPART	0,74	0,71	0,37	0,39	0,66	0,77	0,47	0,55	0,67	0,73	cp=0,02

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

Tabela 8. Características estatísticas de modelos de QSAR para CYP3A4 atribuído para o conjunto teste do conjunto B.

	Conjunto teste (188*)*									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-PLSDA	0,78	0,73	0,28	0,33	0,67	0,80	0,28	0,39	0,61	0,61
MACSS-SVM	0,87	0,86	0,61	0,63	0,85	0,87	0,59	0,69	0,78	0,61
MACCS-KNN	0,79	0,73	0,36	0,38	0,65	0,82	0,38	0,48	0,65	0,61
MACSS-CTREE[1]	0,78	0,73	0,28	0,33	0,67	0,80	0,28	0,39	0,61	0,61
MACSS, CTREE[2]	0,78	0,71	0,36	0,37	0,60	0,82	0,41	0,49	0,66	0,61
MACSS-GBM	0,85	0,81	0,58	0,58	0,75	0,88	0,62	0,68	0,78	0,61
MACCS-RPART	0,78	0,72	0,32	0,35	0,63	0,81	0,34	0,44	0,64	0,61

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

7.1.3 Conjunto C

Tabela 9. Características estatísticas de modelos de QSAR para CYP3A4 avaliado por *5-fold* para o conjunto modelagem do conjunto C.

	Conjunto modelagem (1.139*)*										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACSS-PLSDA	0,81	0,73	0,29	0,33	0,63	0,83	0,28	0,38	0,62	0,71	ncomp=3
MACCS-SVM	0,85	0,80	0,49	0,51	0,74	0,87	0,47	0,58	0,71	0,71	C=4
MACCS-KNN	0,83	0,76	0,41	0,43	0,67	0,85	0,41	0,51	0,68	0,71	K=5
MACCS- CTREE[1]	0,81	0,72	0,37	0,38	0,59	0,85	0,40	0,48	0,66	0,71	mincri=0,459
MACCS-CTREE[2]	0,79	0,68	0,30	0,31	0,51	0,84	0,36	0,42	0,63	0,71	maxd=8
MACSS-GBM	0,85	0,79	0,51	0,51	0,69	0,88	0,53	0,60	0,73	0,71	shrinkage=0,1
MACSS-RPART	0,78	0,67	0,32	0,32	0,48	0,85	0,43	0,46	0,65	0,71	cp=0

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

Tabela 10. Características estatísticas de modelos de QSAR para CYP3A4 atribuídos para o conjunto teste do conjunto C.

	Conjunto teste (284*)*									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-PLSDA	0,83	0,65	0,15	0,18	0,44	0,85	0,14	0,22	0,55	0,59
MACSS-SVM	0,86	0,77	0,39	0,42	0,67	0,88	0,36	0,47	0,66	0,59
MACCS-KNN	0,85	0,73	0,40	0,41	057	0,89	0,43	0,49	0,68	0,59
MACSS-CTREE[1]	0,77	0,55	0,08	0,08	0,25	0,84	0,18	0,21	0,54	0,59
MACSS, CTREE[2]	0,75	0,57	0,15	0,15	0,28	0,86	0,32	0,30	0,58	0,59
MACSS-GBM	0,81	0,66	0,31	0,31	0,43	0,89	0,43	0,43	0,66	0,59
MACCS-RPART	0,79	0,64	0,31	0,31	0,39	0,89	0,50	0,44	0,67	0,59

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

7.1.4 Conjunto D

Tabela 11. Características estatísticas de modelos de QSAR para CYP3A4 avaliado por *5-fold* para o conjunto modelagem do conjunto D.

	Modeling Set (1.520*)*										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACSS-PLSDA	0,86	0,77	0,33	0,37	0,67	0,87	0,28	0,39	0,62	0,72	ncomp=12
MACCS-SVM	0,88	0,81	0,50	0,51	0,71	0,90	0,46	0,56	0,71	0,72	C=16
MACCS-KNN	0,87	0,77	0,44	0,45	0,65	0,89	0,43	0,51	0,69	0,72	K=5
MACCS- CTREE[1]	0,82	0,66	0,29	0,29	0,45	0,88	0,34	0,39	0,63	0,72	mincri=0,378
MACCS-CTREE[2]	0,83	0,68	0,33	0,33	0,47	0,88	0,39	0,43	0,65	0,72	maxd=11
MACSS-GBM	0,87	0,77	0,50	0,51	0,63	0,91	0,54	0,58	0,74	0,72	shrinkage=0,1
MACSS-RPART	0,81	0,66	0,32	0,32	0,44	0,89	0,43	0,43	0,66	0,72	cp=0

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

Tabela 12. Características estatísticas de modelos de QSAR para CYP3A4 atribuídos para o conjunto teste do conjunto D.

	Conjunto teste (378*)*									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-PLSDA	0,86	0,71	0,21	0,25	0,55	0,88	0,18	0,27	0,58	0,64
MACSS-SVM	0,89	0,80	0,42	0,45	0,71	0,90	0,35	0,47	0,66	0,64
MACCS-KNN	0,86	0,69	0,31	0,31	0,48	0,89	0,32	0,39	0,63	0,64
MACSS-CTREE[1]	0,86	0,71	0,29	0,31	0,53	0,89	0,26	0,35	0,61	0,64
MACSS, CTREE[2]	0,87	0,73	0,25	0,29	0,58	0,88	0,21	0,30	0,59	0,64
MACSS-GBM	0,88	0,77	0,45	0,46	0,63	0,91	0,44	0,52	0,70	0,64
MACCS-RPART	0,86	0,69	0,25	0,28	0,50	0,88	0,24	0,32	0,60	0,64

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

7.1.5 Conjunto E

Tabela 13. Características estatísticas de modelos de QSAR para CYP3A4 avaliado por *5-fold* para o conjunto modelagem do conjunto E.

	Conjunto modelagem (1.899*)*										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACSS-PLSDA	0,88	0,77	0,33	0,37	0,65	0,89	0,28	0,39	0,63	0,66	ncomp=16
MACCS-SVM	0,90	0,82	0,49	0,52	0,74	0,91	0,44	0,55	0,70	0,66	C=16
MACCS-KNN	0,88	0,77	0,44	0,45	0,64	0,91	0,41	0,50	0,68	0,66	K=5
MACCS- CTREE[1]	0,87	0,72	0,36	0,37	0,55	0,90	0,35	0,43	0,65	066	mincri=0,092
MACCS-CTREE[2]	0,86	0,72	0,36	0,37	0,54	0,90	0,36	0,43	0,66	0,66	maxd=10
MACSS-GBM	0,89	0,79	0,53	0,54	0,66	0,92	0,53	0,59	0,74	0,66	shrinkage=0,1
MACSS-RPART	0,86	0,70	0,36	0,36	0,49	0,90	0,41	0,45	0,67	0,66	cp=0

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

Tabela 14. Características estatísticas de modelos de QSAR para CYP3A4 atribuídos para o conjunto teste do conjunto E.

	Conjunto teste (474)									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-PLSDA	0,89	0,78	0,20	0,27	0,67	0,90	0,14	0,23	0,56	0,53
MACSS-SVM	0,90	0,81	0,44	0,47	0,69	0,92	0,38	0,49	0,68	0,53
MACCS-KNN	0,90	0,77	0,42	0,43	0,61	0,92	0,38	0,47	0,67	0,53
MACSS-CTREE[1]	0,90	0,88	0,30	0,39	0,86	0,91	0,21	0,33	0,60	0,53
MACSS, CTREE[2]	0,90	0,78	0,37	0,40	0,64	0,92	031	0,42	0,64	0,53
MACSS-GBM	0,88	0,69	0,37	0,37	0,46	0,92	0,41	0,44	0,68	0,53
MACCS-RPART	0,86	0,65	0,30	0,30	0,38	0,92	0,38	0,38	0,65	0,53

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

7.1.6 Conjunto F

Tabela 15. Características estatísticas de modelos de QSAR para CYP3A4 avaliado por *5-fold* para o conjunto modelagem do conjunto F

	Conjunto modelagem (3475*)*										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACSS-PLSDA	0,92	0,83	0,14	0,24	0,75	0,92	0,09	0,16	0,54	0,66	ncomp=25
MACCS-SVM	094	0,88	0,45	0,50	0,83	0,94	0,33	0,48	0,66	0,66	C=8
MACCS-KNN	0,93	0,79	0,40	0,42	0,65	0,94	0,32	0,43	0,65	0,66	K=5
MACCS- CTREE[1]	0,91	0,75	0,16	0,23	0,58	0,92	0,11	0,18	0,55	0,66	mincri=0,092
MACCS-CTREE[2]	0,91	0,73	0,18	0,23	0,54	0,92	0,13	0,21	0,56	0,66	maxd=15
MACSS-GBM	0,93	0,80	0,48	0,49	0,65	0,95	0,42	0,51	0,70	0,66	shrinkage=0,1
MACSS-RPART	0,91	0,71	0,31	0,32	0,49	0,93	0,27	0,35	0,62	0,66	cp=0

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

Tabela 16. Características estatísticas de modelos de QSAR para CYP3A4 atribuídos para o conjunto teste do conjunto F.

	Test Set (868*)*									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-PLSDA	0,93	0,97	0,10	0,22	1,00	0,93	0,05	0,10	0,53	0,58
MACSS-SVM	0,94	0,83	0,37	0,42	0,71	0,94	0,27	0,39	0,63	0,58
MACCS-KNN	0,93	0,75	0,37	0,39	0,55	0,95	0,32	0,41	0,65	0,58
MACSS-CTREE[1]	0,92	0,70	0,21	0,24	0,46	0,94	0,16	0,24	0,57	0,58
MACSS, CTREE[2]	0,92	0,66	0,19	0,21	0,38	0,94	0,16	0,23	0,57	0,58
MACSS-GBM	0,92	0,72	0,37	0,37	0,48	0,95	0,35	0,41	0,66	0,58
MACCS-RPART	0,92	0,69	0,26	0,27	0,44	0,94	0,22	0,29	0,60	0,58

PLS-DA: *Partial least squares discriminant analysis;* SVM: *Support Vector Machine*; SVM: svm radialCost; *k*-NN: *k-Nearest Neighbors*; CTREE: *Conditional Inference Tree;* CTREE[1]*:* Critério mínimo CTREE[2]: Critério máximo; GBM: *Gradient boosting method*; CART: *rpart*; MACCS: *MACCS keys;* CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient*; MCC: Coeficiente de correlação de Matthews; Se: Sensitividade; Sp: Especificidade; F1:*F-score*; AUC: Área sob a curva.

7.1.7 Otimização dos modelos utilizando o conjunto B

Tabela 17. Características estatísticas de modelos de QSAR gerados para substratos de CYP3A4 avaliado por 5-fold para o conjunto modelagem, utilizando o conjunto B.

	Conjunto treinamento (760)										
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**	**Parâmetro do modelo**
MACCS-SVM	0,81	0,79	0,52	0,53	0,75	0,83	0,57	0,64	0,74	0,72	C=2
featMorgan-SVM	0,82	0,81	0,61	0,61	0,79	0,83	0,72	0,75	0,80	0,70	C=4
AtomPair-SVM	0,79	0,79	0,56	0,56	0,75	0,82	0,70	0,73	0,78	0,69	C=4
PubChem-SVM	0,81	0,81	0,57	0,58	0,81	0,81	0,65	0,72	0,78	0,72	C=4
MACCS-PLS	0,78	0,77	0,42	0,45	0,74	0,79	0,45	0,56	0,69	0,72	ncomp=2
featMorgan-PLS	0,73	0,72	0,41	0,42	0,68	0,75	0,57	0,62	0,70	0,70	ncomp=3
AtomPair –PLS	0,75	0,74	0,46	0,47	0,71	0,77	0,61	0,66	0,73	0,69	ncomp=5
PubChem-PLS	0,77	0,79	0,48	0,51	0,82	0,76	0,51	0,63	0,72	0,72	ncomp=5
MACCS-KNN	0,79	0,79	0,44	0,47	0,78	0,79	0,43	0,56	0,69	0,72	k=13
featMorgan-KNN	0,68	0,66	0,31	0,31	0,59	0,73	0,55	0,57	0,65	0,70	k=7
AtomPair –KNN	0,66	0,64	0,27	0,27	0,58	0,70	0,49	0,53	0,63	0,69	k=5
PubChem-KNN	0,74	0,73	0,43	0,44	0,71	0,76	0,53	0,61	0,70	0,72	k=5
MACCS-GBM	0,83	0,80	0,57	0,57	0,75	0,85	0,64	0,69	0,77	0,72	interaction depth= 21 ntrees=50, shrin=0,1
featMorgan-GBM	0,84	0,83	0,65	0,65	0,80	0,86	0,77	0,78	0,82	0,70	interaction depth= 22, ntrees=850, shrink=0,1
AtomPair -GBM	0,80	0,79	0,58	0,58	0,75	0,84	0,75	0,75	0,79	0,69	interaction depth= 21, ntrees=150, shrin=0,1
PubChem-GBM	0,84	0,83	0,65	0,65	0,80	0,86	0,76	0,78	0,82	0,72	Interation depth= 20, ntrees=1000, shrin=0,1

SVM: *Support Vector Machine*; PLS-DA: *PLS Discriminant Analysis*; *k*-NN: *k-Nearest Neighbors*; GBM: *Gradient boosting method*; MACCS: *MACCS keys*; PubChem: *PubChem Fingerprints*; FeatMorgan: *Circular fingerprint based on the Morgan algorithm and feature invariants (FCFP-like);* Atom Pair: *Atom pair fingerprints*; CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient;* MCC: Coeficiente de correlação de Matthews; Se: Sensibilidade; Sp: Especificidade; *F1:F-score;* AUC: Área sob a curva.

Tabela 18. Características estatísticas dos modelos de QSAR para CYP3A4 atribuídos para o conjunto teste.

	Conjunto teste (188)									
Nome do modelo	**Acurácia**	**CCR**	**Kappa**	**MCC**	**Se**	**Sp**	**Precisão**	**F1**	**AUC**	**Cobertura**
MACCS-SVM	0,79	0,76	0,38	0,41	0,73	0,80	0,37	0,49	0,66	0,82
featMorgan-SVM	0,71	0,69	0,39	0,39	0,58	0,80	0,64	0,61	0,70	0,85
AtomPair-SVM	0,77	0,75	0,50	0,50	0,68	0,82	0,67	0,68	0,75	0,89
PubChem-SVM	0,75	0,73	0,44	0,44	0,66	0,79	0,59	0,62	0,71	0,85
MACCS-PLSDA	0,74	0,69	0,18	0,23	0,62	0,76	0,19	0,29	0,57	0,82
FeatMorgan-PLSDA	0,71	0,68	0,33	0,34	0,62	0,75	0,46	0,53	0,66	0,85
AtomPair –PLSDA	0,74	0,71	0,41	0,41	0,65	0,78	0,57	0,61	0,70	0,89
PubChem-PLSDA	0,71	0,69	0,31	0,33	0,65	0,73	0,39	0,49	0,64	0,85
MACCS-KNN	0,79	0,77	0,37	0,41	0,75	0,79	0,35	0,48	0,65	0,82
featMorgan-KNN	0,71	0,69	0,39	0,39	0,58	0,80	0,64	0,61	0,70	0,85
AtomPair –KNN	0,69	0,67	0,34	0,34	0,57	0,77	0,59	0,58	0,67	0,89
PubChem-KNN	0,67	0,65	0,31	0,31	0,52	0,78	0,63	0,57	0,66	0,85
MACCS-GBM	0,75	0,68	0,34	0,34	0,56	0,81	0,47	0,51	0,66	0,82
featMorgan-GBM	0,70	0,68	0,36	0,36	0,55	0,80	0,66	0,60	0,69	0,85
AtomPair -GBM	0,76	0,75	0,50	0,50	0,66	0,83	0,70	0,68	0,75	0,89
PubChem-GBM	0,73	0,71	0,42	0,42	0,61	0,81	0,66	0,63	0,72	0,85

SVM: *Support Vector Machine* ; PLS-DA: *PLS Discriminant Analysis*; *k*-NN: *k-Nearest Neighbors*; GBM: *Gradient boosting method*; MACCS: *MACCS keys*; PubChem: *PubChem Fingerprints*; FeatMorgan: *Circular fingerprint based on the Morgan algorithm and feature invariants (FCFP-like);* Atom Pair: *Atom pair fingerprints*; CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient;*MCC: Coeficiente de correlação de Matthews; Se: Sensibilidade; Sp: Especificidade; *F1:F-score;* AUC: Área sob a curva.

Tabela 19. Resultados estatísticos para modelos de QSAR para CYP3A4 avaliados pelo método de randomização de Y.

Nome do modelo	Acurácia	CCR	Kappa	Se	Sp	AUC
MACCS-SVM	0,54±0,03	0,53±0,12	0,06±0,06	0,30±0,12	0,73±0,12	0,54±0,04
featMorgan-SVM	0,53±0,03	0,52±0,13	0,04±0,06	0,28±0,13	0,77±0,13	0,54±0,03
AtomPair-SVM	0,53±0,03	0,53±0,12	0,05±0,06	0,30±0,13	0,75±0,11	0,54±0,04
PubChem-SVM	0,53±0,03	0,52±0,15	0,04±0,06	0,29±0,15	0,75±0,15	0,54±0,03
MACCS-PLSDA	0,53±0,03	0,53±0,06	0,05±0,07	0,48±0,06	0,57±0,06	0,54±0,03
featMorgan-PLSDA	0,53±0,04	0,53±0,06	0,06±0,09	0,51±0,06	0,55±0,06	0,54±0,03
AtomPair –PLSDA	0,53±0,04	0,53±0,05	0,06±0,07	0,51±0,05	0,54±0,06	0,54±0,03
PubChem-PLSDA	0,53±0,03	0,53±0,06	0,06±0,07	0,49±0,06	0,57±0,05	0,53±0,03
MACCS-KNN	0,52±0,04	0,52±0,06	0,05±0,07	0,47±0,06	0,58±0,06	0,53±0,04
featMorgan-KNN	0,53±0,04	0,52±0,08	0,05±0,07	0,50±0,07	0,55±0,08	0,54±0,03
AtomPair –KNN	0,53±0,04	0,53±0,07	0,05±0,07	0,50±0,07	0,55±0,07	0,53±0,03
PubChem-KNN	0,53±0,04	0,55±0,06	0,06±0,08	0,49±0,07	0,57±0,06	0,54±0,04
MACCS-GBM	0,54±0,03	0,32±0,05	0,08±0,06	0,52±0,05	0,56±0,05	0,54±0,03
featMorgan-GBM	0,54±0,03	0,32±0,05	0,07±0,07	0,51±0,05	0,56±0,06	0,54±0,03
AtomPair –GBM	0,54±0,04	0,32±0,06	0,09±0,08	0,51±0,07	0,57±0,05	0,54±0,03
PubChem-GBM	0,54±0,04	0,32±0,06	0,07±0,08	0,52±0,06	0,56±0,06	0,54±0,04

SVM: *Support Vector Machine* ; PLS-DA: *PLS Discriminant Analysis*; *k*-NN: *k-Nearest Neighbors*; GBM: *Gradient boosting method*; MACCS: *MACCS keys*; PubChem: *PubChem Fingerprints*; FeatMorgan: *Circular fingerprint based on the Morgan algorithm and feature invariants (FCFP-like);* Atom Pair: *Atom pair fingerprints*; CCR: Acurácia balanceada; Kappa: *Cohen's kappa coefficient;* MCC: Coeficiente de correlação de Matthews; Se: Sensibilidade; Sp: Especificidade; *F1:F-score;* AUC: Área sob a curva.

7.1.8 Resultados estatísticos para modelos de inibidores de CYP3A4.

Tabela 20. Resultados estatísticos dos modelos de QSAR binário e multiclasse para inibidores de CYP3A4 avaliados por *5-fold*.

Nome do modelo	Acurácia	Precisão	Sensibilidade	F1
Morgam-SVM Binário	0,76	0,77	0,78	0,74
Morgam-SVM Multiclasse	0,66	0,66	0,66	0,66
Morgan RF-Binário	0,76	0,78	0,76	0,77
Morgan RF-Multiclasse	0,69	0,69	0,69	0,68

SVM: *Support Vector Machine*; RF: *Radon Forest*; Kappa: *Cohen's kappa coefficient;* Se: Sensibilidade; F1*: F-score.*