UNIVERSIDADE DE SÃO PAULO
FACULDADE DE CIÊNCIAS FARMACÊUTICAS
CURSO DE GRADUAÇÃO EM FARMÁCIA-BIOQUÍMICA

Estudos de Relação Quantitativa entre Estrutura-Atividade de uma série de aminas heterocíclicas com atividade antidepressiva *in vitro*

Gustavo Henrique Marques Sousa

Trabalho de Conclusão do Curso de Farmácia-Bioquímica da Faculdade de Ciências Farmacêuticas da Universidade de São Paulo.

Orientador: Prof. Dr. Gustavo Henrique Goulart Trossini

São Paulo

2020

SUMÁRIO

LISTA DE ABREVIATURAS

OMS	Organização Mundial de Saúde
5-HT	5-Hidroxitriptamina
MAOIs	*Monoamine Oxidase Inhibitor*
TRI	*Triple Reuptake Inhibitor*
SERT	*Serotonine Transporter*
DAT	*Dopamine Transporter*
NET	*Norepinefrine Transporter*
HTS	*High Througuput Screening*
QSAR	*Quantitative Structure Activity Relationship*
QSPR	*Quantitative Structure-Property Relationship*
ML	*Machine Learning*
ANN	*Artificial Neural Network*
PLS	*Partial Least Squares*
MLR	*Multiple Linear Regression*
SVM	*Support Vector Machine*
RF	*Random Forest*
RMSE	*Root Mean Square Error*
MAE	*Mean Square Error*
CCC	*Concordance Correlation Coeficient*

RESUMO

SOUSA, G. H. M. **Estudos de Relação Quantitativa entre Estrutura-Atividade de uma série de aminas heterocíclicas com atividade antidepressiva in vitro**. 2020. no. f. Trabalho de Conclusão de Curso de Farmácia-Bioquímica – Faculdade de Ciências Farmacêuticas – Universidade de São Paulo, São Paulo, 2020.

Palavras-chave: Planejamento de fármacos, Química Farmacêutica, *Machine Learning*, QSAR

INTRODUÇÃO: A depressão é uma desordem do sistema nervoso central e que acomete indivíduos no mundo todo. O arsenal terapêutico atual possui características indesejadas, como a demora do tempo de ação além de diversos efeitos adversos. Nesse contexto, a quimioinformática é uma ciência que pode ser aplicada no descobrimento e otimização de moléculas promissoras no tratamento da depressão e de diversas outras doenças, se destacando metodologias como a Relação Quantitativa entre Estrutura e Atividade (QSAR). **OBJETIVO**: Explorar técnicas suportadas pela literatura e que visam a obtenção de modelos com capacidade de predição da atividade biológica, neste caso atividade antidepressiva observada *in vitro*, baseados na estrutura química dos compostos, assim como explorar suas respectivas vantagens e desvantagens. **MATERIAL E MÉTODOS**: A partir de compostos obtidos na literatura com atividade antidepressiva *in* vitro, utilizou-se um fluxograma típico de QSAR, com partição randômica da série teste e treino, validação interna e externa dos dados e métricas pertinentes nas avaliações das respectivas performances. Foram priorizados *softwares* gratuitos para a representação das estruturas químicas, cálculo de descritores para o todo processo de modelagem, garantindo fácil acesso à reprodutibilidade dos resultados. Para os modelos de predição de atividade, foram empregados métodos lineares e não-lineares, explorando alguns dos algoritmos utilizados na literatura como *PLS*, *Random Forest* e *SVM,* utilizando para tal finalidade a linguagem de programação R. Não obstante, os parâmetros farmacocinéticos foram explorados e preditos com auxílio da plataforma *SwissADME*. **RESULTADOS**: Os modelos não-lineares obtidos por *Random Forest* e *SVM* apresentaram performance bastante superior quando comparados aos modelos lineares, com destaque àqueles que foram construídos a partir de descritores mais simples e com significado químico de fácil interpretação. As métricas obtidas para os modelos relacionados à inibição dos transportadores de norepinefrina e dopamina se apresentaram adequadas para um estudo de QSAR. As propriedades farmacocinéticas preditas para as moléculas presentes nesse estudo se apresentaram promissoras no que diz respeito ao desenvolvimento de novos candidatos à fármacos. **CONCLUSÃO**: Entre os modelos não-lineares de dopamina e norepinefrina, aqueles obtidos por RF possuem descritores convergentes, sendo destaque descritores tridimensionais (RDF), tamanho de cadeia principal e estado eletrônico de nitrogênio presente na estrutura, além de apresentarem como característica uma baixa demanda computacional. Sugere-se que para atividade relacionada à serotonina, outras técnicas sejam empregadas, visto que neste estudo não foi possível obter um modelo adequado para predição da inibição deste receptor.

INTRODUÇÃO

1.1 Epidemiologia da Depressão

De acordo com a Manual Estatístico e Diagnóstico de Desordens Mentais (DSM–V; 5ª edição, *American Psychiatric Association*, 2013), o transtorno depressivo maior, conhecida popularmente como depressão, é caracterizado por uma série de sintomas dentre eles: perda de interesse em diversas atividades cotidianas; alterações no sono e em atividades psicomotoras; sentimento de culpa; dificuldade em se concentrar assim como na tomada de decisões; pensamentos recorrentes relacionados a morte e suicídio, entre outros sintomas correlatos (PEREZ-CABALERO et al, 2019).

Dados da Organização Mundial de Saúde mostram que a depressão é responsável por 10% das doenças não fatais e é, do ponto de vista global, a maior responsável por anos perdidos de trabalho por invalidez, do que qualquer outra condição ou enfermidade, mostrando seu potencial incapacitante para com o indivíduo acometido. (OMS, 2016). Estudos epidemiológicos apontam para uma diferença na prevalência, incidência e morbidade de depressão em mulheres, gênero que possui maiores taxas deste transtorno ao longo da vida adulta, ao contrário da taxa observada em homens que é mais preponderante durante a adolescência. Contudo, não há conclusões definitivas sobre as diferenças de gênero e incidência de depressão (PICCINELLI, WILKINSON, 2000). Ainda que seja menos prevalente em períodos avançados, a doença ao acometer idosos pode representar graves consequências, uma vez que a taxa de suicídio é maior nesta faixa etária. Apesar de se apresentar em declínio, a taxa de suicídio é maior na população idosa quando comparada a população jovem. Sugere-se que esse fato pode estar relacionado à prevalência de depressão neste grupo (FISKE, WETHERRELL, 2009).

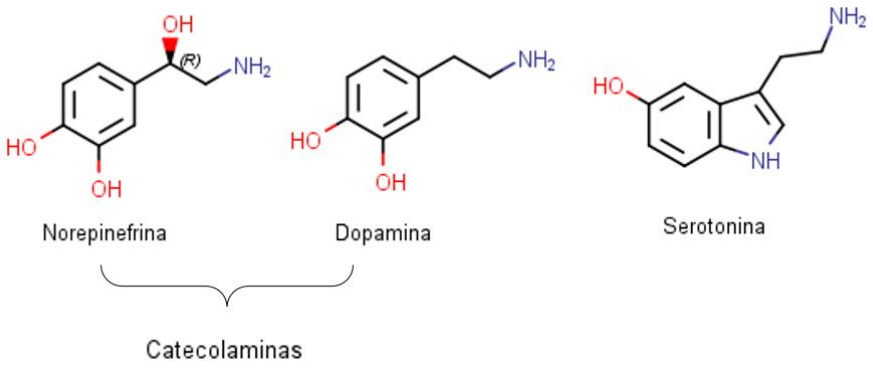
Dada a relevância epidemiológica do transtorno depressivo, bem como o agravante fator incapacitante da doença, é preciso considerar veementemente a inclusão da depressão como uma prioridade na saúde pública mundial nos próximos anos (FERRARI et al, 2013).

1.2 Fisiopatologia e Tratamento da Depressão

Estudos sobre a neurobiologia e origem deste transtorno foram iniciados em meados dos anos 50. Estes surgiram da observação de pacientes com tuberculose que ao serem tratados com iproniazida apresentavam melhora no humor, visto que alguns destes indivíduos também sofriam de depressão (LOOMER et al, 1957).

Ao longo dos anos conseguintes, foi possível observar o efeito de alguns fármacos com propriedade de modificar a biodisponibilidade de catecolaminas (i.e., norepinefrina e dopamina) e que melhoravam os sintomas depressivos, e que por consequência deram origem a hipótese catecolaminérgica (SCHILDKRAUT, 1965).

Figura 1 - Estruturas químicas das monoaminas: norepinefrina, dopamina e serotonina.



Fonte: PUBCHEM (CID - 439260, 681, 5202), acessado em 10 de setembro de 2020

Alguns anos depois, reconheceu-se o papel da serotonina (5-HT) na regulação do transtorno depressivo, subsidiando a hipótese monoaminérgica, amplamente conhecida atualmente. Assim, a terapêutica antidepressiva foi suportada por meio de fármacos que possuíam como efeito resultante comum o aumento da disponibilidade das monoaminas na fenda sináptica, como os Antidepressivos Tricíclicos (TCAs) e Inibidores da enzima Monoamina Oxidase (MAOIs), sendo por muitos anos a linha de frente no tratamento deste transtorno (PEREZ-CABALERO et al, 2019).

Durante o desenvolvimento de terapias antidepressivas, moléculas com inibição específica de serotonina e norepinefrina tomaram a liderança por possuírem um perfil mais seguro que antidepressivos tricíclicos e inibidores de monoamina oxidase. Apesar da maior tolerabilidade, há possibilidades significativas de melhora no tempo de ação inicial desses fármacos, já que a redução dos sintomas e do quadro clínico da depressão se dá em um período aproximado de 2-4 semanas (MARKS et al., 2008).

Uma outra alternativa, ainda considerando a hipótese monoaminérgica, é a utilização dos Inibidores de Recaptura Tripla ou TRIs *(*do inglês *triple-reuptake-inhibitors*) que possuem a capacidade de inibir simultaneamente os transportadores de norepinefrina, dopamina e serotonina. A hipótese assume que o TRI estaria ligado a ativação sistema de recompensa dopaminérgico e possibilitaria reduzir efeitos colaterais que outras classes não são capazes, dada especificidade destas por apenas um alvo (LANE, 2014). Entretanto, não há consenso na literatura se a inibição dos 3 principais transportadores responsáveis pela recaptura de monoaminas - Transportador de Serotonina (SERT), Transportador de Dopamina (DAT) e Transportador de Norepinefrina (NET) - levaria a uma resposta antidepressiva mais rápida e mais eficaz. É sabido, todavia, que as moléculas de inibição tripla possuem um potencial de atuar no tratamento da depressão por um espectro maior dos sintomas como por exemplo, a anedonia, reduzindo os efeitos colaterais resultantes dos tratamentos convencionais atualmente empregados. Ainda assim, é um desafio do ponto de vista de planejamento de fármacos obter uma molécula que apresente, concomitantemente, um perfil de inibição significativo para os três transportadores, com boa biodisponibilidade oral e não menos importante, poucos efeitos adversos (SHARMA et al, 2015).

1.3 QSAR e Quimioinformática

A quimioinformática é um campo da ciência que combina elementos da química, biologia e ciência da computação para transformar dados químicos e biológicos em conhecimentos úteis que suportam a tomada de decisão no planejamento e otimização de fármacos (CHEN et al, 2018).

Nesse contexto, a triagem de alto desempenho, do inglês *high-throughput screening* (HTS) é amplamente utilizada pela indústria farmacêutica na busca de novos compostos biologicamente ativos. É uma técnica robotizada e automatizada de ensaios experimentais em grande escala e que possui uma taxa relativamente baixa de descoberta de compostos líderes, quando comparada à triagem virtual. Esta, por sua vez, enquanto oferece boas taxas na descoberta de novos medicamentos gera uma imensa quantidade de dados químicos. Já triagem exclusivamente virtual de compostos, ou seja, sem o uso de reagentes e emprego de síntese em alta escala, possui usualmente taxa de descoberta de novos compostos biologicamente ativos variando de 1% a 40% num modelo robusto e bem validado. Seu custo é significativamente menor que o HTS, por se tratar de um método computacional e não experimental (NEVES et al, 2018).

A quimioinformática é um ramo da química que tem como foco a transformação de dados químicos, muitas vezes obtidos por HTS, em informações úteis para o planejamento de novos fármacos, empregando técnicas computacionais, matemáticas e estatísticas. Suas aplicações mais práticas estão nas ciências ambientais, saúde, toxicologia e planejamento de fármacos. Entre as técnicas que auxiliaram no desenvolvimento e consolidação deste ramo da ciência, está a Análise Quantitativa da Relação Estrutura-Atividade, do inglês *Quantitative Structure-Activity Relationshi*p (QSAR) que tem como objetivo estabelecer um modelo matemático, baseado em descritores físico-químicos de uma série de moléculas estruturalmente similares e com atividade biológica definida (ALVES, V. et al, 2017). É comum encontrar na literatura o uso dos prefixos “q” e “Q” no acrônimo SAR, indicando modelos que abordam problemas qualitativos e quantitativos, respectivamente (BURBIDGE et al., 2001).

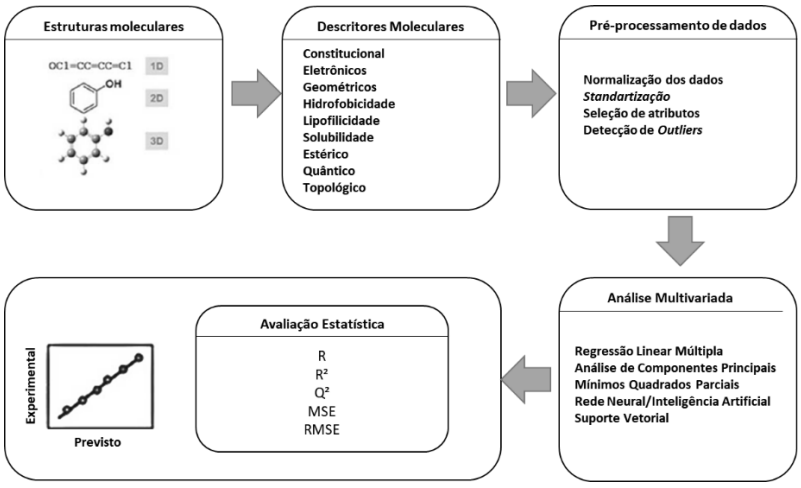
A técnica foi desenvolvida há mais de 50 anos, por HANSCH e FUJITA, (1964) e desde então, diferentes abordagens - QSAR-2D, QSAR-3D, Hologram-QSAR, Fragment-QSAR - aliadas a uma gama de ferramentas computacionais, que utilizam *Machine Learning* em sua estrutura sistemática, podem ser utilizadas para se obter modelos preditivos no descobrimento de novas moléculas biologicamente ativas. Tal ferramenta é muito importante do ponto de vista acadêmico e industrial, visto que os modelos gerados podem fornecer

indicativos relevantes antes mesmo da síntese e avaliação biológica experimental, reduzindo o tempo no ciclo de desenvolvimento bem como os custos associado a descoberta de um novo fármaco (NEVES et al., 2018). Diversas abordagens de modelagem por meio da técnica de QSAR são suportadas por técnicas advindas da Estatística e Aprendizado de Máquina (*Machine Learning*). Entre alguns exemplos de algoritmos estão: Árvore de Decisão, Redes Neurais Artificiais, Quadrados Mínimos Parciais, k-Vizinho-Mais-Próximo, Regressão Linear Múltipla, Análise Discriminante e Máquinas de Vetores de Suporte (SVETNIK et. al, 2003). Essas técnicas são capazes de modelar dados mais complexos e não-lineares utilizando para este fim algoritmos que detectam os padrões moleculares relacionados a atividade biológica (NEVES et al., 2020).

Diversas publicações utilizando QSAR são caracterizadas pelo uso de uma pletora de descritores químicos, que são utilizados na construção de modelos lineares, como aqueles foram conduzidos na parte inicial deste trabalho, ou ainda de modelos que relacionem de forma não-linear os descritores e atividade biológica. Esses modelos seguem metodologia rigorosa de validação interna e externa, visando-se obter resultados estatisticamente apropriados e significativamente preditivos (TROPSHA; GOLBRAIKH, 2007).

A figura (2) apresenta um esquema adaptado que mostra de maneira simplificada o processo de obtenção de um modelo QSAR.

Figura 2 - Esquema geral do processo de QSAR



Fonte: DAMALE et. al, 2014 (Adaptado).

Entre as métricas utilizadas para a avaliação de um modelo QSAR estão, entre outras, o Coeficiente de Correlação Linear (R^2), a Raiz do Quadrado Médio do Erro (*RMSE*) e a Média Absoluta do Erro (*MAE*). O R^2 quantifica a preditividade do modelo, ou seja, o quanto ele se mostra eficaz em capturar e prever os dados experimentais e seu cálculo é feito a partir da equação (1). Quanto mais próximo de 1 o R^2 entre os dados experimentais e os dados previstos por um modelo, maior a capacidade do modelo em "explicar" estes respectivos dados experimentais. Já o *RMSE* disposto na equação (2), do inglês *Root Mean Square Error* está relacionado à comparação da performance em dados externos e internos, bem como entre modelos obtidos por diferentes metodologias. O *RMSE* também pode ser interpretado como o desvio padrão do resíduo, que é a diferença entre o valor experimental e o valor estimado. Por fim, a *MAE* do inglês *Mean Absolute Erro*r é bastante encontrada na literatura e representa a média absoluta do resíduo. Apresenta a vantagem de estar na mesma unidade de medida dos valores da variável resposta, calculado a partir da equação (3). Todas essas equações apresentam argumentos nos quais: y_{imean} é a média correspondente; y_{iexp} é o valor experimental correspondente; y_{ipred} é o valor predito correspondente e n o número de amostras utilizados no modelo (FERREIRA, 2002).

$$(1) \qquad R^2 = 1 - \frac{\sum(y_{iexp} - y_{ipred})^2}{\sum(y_{iexp} - y_{imean})^2}$$

$$(2) \qquad RMSE = \sqrt{\frac{\sum(y_{iexp} - y_{ipred})^2}{n}}$$

$$(3) \qquad MAE = \left|\frac{\sum(y_{iexp} - y_{ipred})}{n}\right|$$

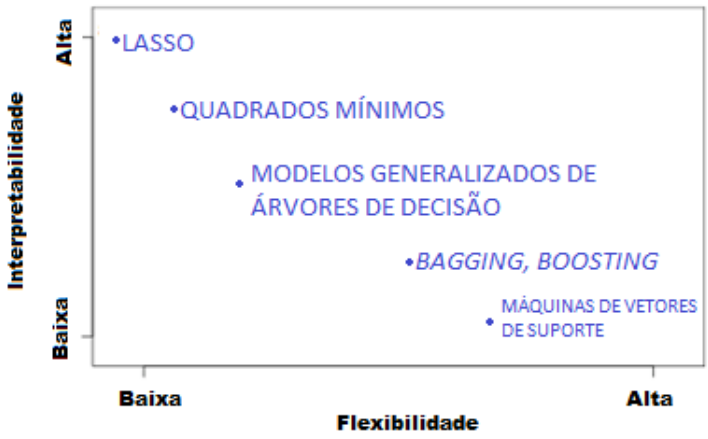
O principal propósito da validação é fornecer um modelo estatisticamente significativo, como uma consequência adequada de causa e efeito e evitar uma relação numérica obtida ao acaso (KIRALJ e FERREIRA, 2009). Nesse sentido, a validação externa tem sido um extenso debate entre aqueles que utilizam QSAR já que alguns autores discutem que a capacidade preditiva "real" de um modelo só pode ser adequadamente estimada utilizando um grupo de teste externo, ou seja, compostos que nunca foram utilizados na construção dos

modelos. Ainda assim, a validação externa pode conter algumas desvantagens quando o grupo teste utilizado possui alguma distribuição em particular ou diferirem estruturalmente do grupo utilizado para treinar o modelo (GRAMATICA; SANGION, 2020).

1.4 Conceitos iniciais de métodos não lineares

Nos métodos de Aprendizado de Máquina, há uma situação de perda ou ganho de informações com relação a interpretabilidade do modelo em função de sua flexibilidade na explicação da resposta. Por exemplo, do ponto de vista de capacidade preditiva da resposta, um modelo linear que é muito menos flexível que um modelo de árvore de decisão, é mais interpretativo que o último, e por se tratar de um modelo linear é possível ainda comparar a influência e intensidade das diversas variáveis dependentes em relação a variável resposta. Portanto, podemos afirmar que para problemas inferenciais, na qual se deseja explicar a influência de determinados parâmetros em uma resposta, um modelo menos flexível é indicado. Porém, se o que se deseja é apenas prever a resposta, um modelo com alta flexibilidade pode ser a melhor opção (JAMES et. al, 2013).

Figura 3 - Uma representação visual da relação entre métodos flexíveis e sua capacidade de interpretação.



Fonte: JAMES, et. al, 2013 (adaptado)

Algumas metodologias utilizadas em quimioinformática são classificados como “métodos caixa preta”, derivados da nomenclatura em inglês “*Black Box*

Methods". Apesar de sua excelente capacidade de predição, é incapaz de oferecer recursos de como e por que as variáveis independentes, nesse caso os descritores calculados, estão relacionadas com as respostas e este é o principal motivo de um método ser classificado como tal. No caso de modelos mais simples como regressões multilineares, a influência e contribuição de descritores é mais clara e abre espaço para a interpretação química ou mecanística. Outros métodos como a Floresta Aleatória, do inglês "*Random Forest*", permitem que essa informação seja inferida por meio da importância atribuída a cada variável pelo próprio algoritmo (MITCHELL, 2014). Métodos de Aprendizagem de Máquina, mais especificamente os chamados de Conjunto de Árvores podem ser uma escolha adequada para modelagem QSAR, pois há uma combinação de propriedades desejáveis do método de Árvore de Decisão Simples aliado a uma alta performance na predição da variável resposta (SVETINIK et al., 2004). O algoritmo *Random Forest*, é um dos métodos contemplados neste trabalho (BREIMAN, 2001).

1.4.1 Algoritmo *Random Forest*

Random Forest é um algoritmo utilizado em problemas de classificação ou regressão, baseado em árvores de decisão. Cada árvore de decisão é criada utilizando *bootsrap* (um método de amostragem randômica) das amostras dos dados de treinamento e seleção aleatória de variáveis na construção de cada árvore de decisão. As predições são feitas pela maioria absoluta dos votos em determinada classe, no caso de uma classificação, ou na média das estimativas no caso de uma regressão. A capacidade preditiva e informativa deste método já foi bem estabelecida em QSAR, visto sua capacidade em lidar com a alta dimensionalidade dos dados, aliada à sua robustez a variáveis irrelevantes na resposta e também por oferecer possível interpretação do modelo construído. O treino do algoritmo se dá pelo método de *bootstrap* das amostras, ou seja, retira-se aleatoriamente do grupo treino um subconjunto de *n* amostras. Para cada grupo de amostras retirada, uma Árvore de Decisão é gerada, e em cada nó de decisão seleciona-se aleatoriamente um grupo de descritores definido pelo argumento "*Mtry*", que se traduz como o subconjunto de descritores que serão testados em cada nó de decisão. Quando "*Mtry*" é numericamente igual ao

número de descritores totais, tem-se um caso especial em que o algoritmo *Random Forest* é igual ao algoritmo *Bagging.* As Árvores de Decisão são geradas até atingir seu tamanho máximo. As etapas anteriores são repetidas e, no caso de uma regressão, a estimativa para a variável dependente (neste caso a atividade biológica) é uma média de todas as árvores de decisão geradas. (SVETNIK et. al., 2003).

A importância de uma variável num modelo obtido por *Random Forest* é calculada mediante ao "Erro Fora da Bolsa", do inglês "*Out of The Bag Error*." O subconjunto de amostras retiradas do grupo treino é utilizado para a construção da árvore em si e o subconjunto do grupo treino restante é utilizado para estimar o erro de predição, algo semelhante a validação interna, porém a um custo computacional menor. Primeiro é calculada a taxa de Erro Fora da Bolsa para cada árvore e em seguida computa-se a mesma taxa de erro, porém com uma variável permutada. A diferença entre os erros na presença e ausência de uma determinada variável (ou descritor), nos fornece a importância da variável na estimativa da atividade biológica no modelo QSAR. Essa importância atribuida pode ser usada para selecionar e filtrar inicialmente os descritores mais importantes, uma vez que é muito comum em QSAR trabalhar com um número de descritores maior que o número de amostras/estruturas, técnica conhecida como "*Wrapper*", empregada na redução do número de descritores utilizados na obtenção dos modelos (SVETNIK et al., 2004).

Sabendo da moderada ou ainda fraca relação linear entre as variáveis resposta e os descritores, lançou-se mão desta técnica na obtenção de um modelo que fosse adequadamente preditivo e que, dada a natureza da mesma, permita maior interpretabilidade química e ou mecanística.

1.4.2 Algoritmo *Support Vector Machine*

SVM, do inglês "*Support Vector Machine"* é um algoritmo utilizado em diversas aplicações na predição ou classificação de dados, projetando-os num sistema multidimensional, obtido a partir uma função "Kernel" que é, essencialmente, uma função não-linear (ENGEL; GASTEIGER, 2018). O uso de SVM na descoberta de fármacos tem sido cada vez mais frequente, dada sua robustez e capacidade de aplicação para problemas de regressão, classificação

e até triagem/busca de compostos que apresentam diferença estrutural, porém atividades biológicas semelhantes. Conforme descrito anteriormente, num problema de classificação, por exemplo, a ideia central desse algoritmo é a obtenção de uma “regra” de separação para duas ou mais classes de objetos em um espaço multi-dimensional, gerando um hiperplano que é capaz de distinguir os objetos das diferentes classes. Modelos obtidos a partir desse algoritmo são geralmente muito complexos, porém possuem um grande potencial de generalização em dados externos (HEIKAMP; BAJORATH, 2013).

1. OBJETIVOS

2.1. Geral

Este trabalho tem como objetivo explorar técnicas empregadas na literatura que visam a obtenção de modelos com capacidade de predição da atividade biológica baseados na estrutura química dos compostos (QSAR), bem como explorar suas diferenças e vantagens.

2.2. Específico

Aplicar as técnicas exploradas em uma série de aminas heterocíclicas com capacidade antidepressiva observadas *in vitro*, a fim de se obter modelos validados e com capacidade de predição da atividade biológica, levando em consideração vantagens e desvantagens específicas de cada método aplicado.

2. MATERIAL E MÉTODOS

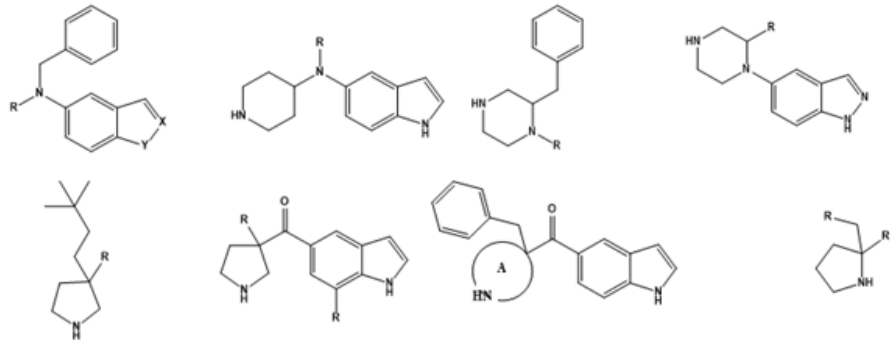
3.1. Softwares

O *software* KNIME foi utilizado para a conversão das moléculas em formatos “SMILES” e “MOLFile”, e apresenta vantagem em sua característica visual e intuitiva para o usuário. Os modelos e gráficos foram obtidos utilizando a linguagem de programação R.

Os compostos utilizados nos estudos de QSAR foram extraídos da literatura, formando um *dataset* final de 77 compostos, retirados de publicações realizadas entre os anos de 2009 e 2010. É importante ressaltar que tais artigos

foram desenvolvidos pelo mesmo grupo de pesquisa da empresa Roche de Palo Alto - USA, garantindo homogeneidade metodológica. Os dados biológicos disponíveis para estes compostos foram obtidos da mesma forma e nas mesmas condições, corroborando com os requisitos para estudos de QSAR. Para estas moléculas, a atividade antidepressiva *in vitro* foi avaliada contra os transportadores de dopamina, noreprinefrina e serotonina (LUCAS et al., 2009; CARTER et. al 2010; LUCAS et al., 2010).

Figura 4 - Formas estruturais gerais das moléculas com atividade antidepressiva, totalizando 77 moléculas;



Legenda: X e Y são átomos diferentes de carbono e oxigênio; A representa um sistema aromático com n átomos. Fonte: Elaborado pelo autor com base nas estruturas presentes em LUCAS et al., 2009; CARTER et. al 2010; LUCAS et al., 2010.

As estruturas foram desenhadas pelo *software* MarvinSketch®, e salvas em formato ".mol", compatível com *softwares* que calculam descritores físico-químicos (MARVIN, 2020).

Para o cálculo dos descritores, utilizou-se o software gratuito Padel-Descriptor, que atualmente é capaz de calcular 1875 descritores (1D, 2D e 3D) e *fingerprints* (FP). Dentre os descritores calculados pelo *software* estão descritores topológicos, eletrônicos, estéricos e quânticos (YAP, 2010).

3.2. Análise da estrutura dos dados

Inicialmente, os dados da atividade biológica e sua correlação linear com a matriz de descritores foram analisados, visando uma compreensão adequada

sobre a estrutura dos dados obtidos e possíveis relações entre si. Os gráficos, assim como os modelos lineares e não-lineares, foram obtidos utilizando *scripts* em R. Os pacotes em R utilizados foram:

- Readxl: Leitura dos dados em .xlsx (MS Excel®);
- pls: Obtenção dos modelos lineares por mínimos quadrados parciais;
- Tidyverse: Manipulação dos dados e gráficos;
- Tidymodels: Obtenção dos modelos não-lineares;

3.3. Análise e predição de propriedades farmacocinéticas

As moléculas foram avaliadas do ponto de vista farmacocinético utilizando a ferramenta gratuita SwissADME, disponível em: <http://www.swissadme.ch>. Esta plataforma possibilita o cálculo e predição de parâmetros-chave na avaliação farmacocinética (absorção, distribuição, metabolismo e excreção) de um composto, além de parâmetros como facilidade sintética, e classe biofarmacêutica (DAINA et al., 2017).

3.4. Modelos preditivos

3.4.1 Grupo teste e grupo treino

A série de 77 compostos foi dividida na razão de 80% (63 estruturas) para o grupo treino e 20% (14 estruturas) para o grupo teste, com a composição dos respectivos grupos realizada de forma aleatória. A fim de se obter reprodutibilidade dos resultados, a linguagem R permite o uso de "*seeds*", que são iniciadores do sistema pseudo-randômico incorporado na linguagem. Foram mantidas as mesmas *seeds* para os modelos lineares e não-lineares relativo a cada transportador, com o objetivo de se comparar a capacidade preditiva obtida utilizando técnicas diferentes, porém utilizando as mesmas estruturas do grupo treino/teste.

3.4.2 Construção de modelos de QSAR

3.4.2.1 Modelos lineares

Para os modelos lineares, utilizou-se o método estatístico de correlação dos mínimos quadrados parciais, do inglês "*Partial Least Squares"* (PLS), na qual foi assumida uma relação linear entre a matriz de descritores e a atividade biológica para cada transportador.

Primeiramente, os descritores foram selecionados com base em sua variância e aqueles que apresentavam um valor próximo a 0 foram descartados, uma vez que não representam nenhuma variação concomitante à resposta biológica, e portanto, nenhuma informação possivelmente ligada à variável dependente. Os descritores que restaram foram selecionados com base em sua correlação linear com as respectivas atividades biológicas, a partir de uma matriz de correlação de Pearson. Pode-se afirmar que modelos com menos descritores são mais facilmente interpretáveis, providenciam melhor performance em amostras que não foram utilizadas para o treinamento e diminuem o risco de *overfit*, uma condição na qual o modelo é incapaz de performar em um grupo externo (GOODARZI et. al, 2012). Para a construção do modelo objetivou-se manter no máximo 10-15 descritores tendo em vista o número de compostos (63) no grupo treino (TOPLISS e COSTELlO, 1972). Dada suas diferenças de dimensionalidade e ordem de grandeza, os descritores foram normalizados, isto é, centrados na média e recalculados para se obter uma variância igual a 1.

3.4.2.2 Modelos não-lineares

Usando a mesma matriz de descritores empregada nos modelos lineares, utilizou-se o algoritmo *Random Forest* na construção de modelos de predição da atividade biológica de inibição dos transportadores de dopamina, norepinefrina e serotonina. Todos os modelos foram ajustados perante validação interna e confirmados utilizando validação externa.

Inicialmente, os descritores foram filtrados e aqueles que apresentavam variância próxima de 0 foram automaticamente retirados da matriz de dados. Nesta etapa, restaram 711 descritores. Todos os descritores foram normalizados, isto é, centrados na média e com variância 1. Com objetivo de

evitar multicolinearidade, descritores com alta correlação linear entre si foram descartados.

Uma vez obtido um modelo com todos os descritores, prosseguiu-se para a construção de um modelo utilizando apenas descritores facilmente interpretáveis. 164 descritores foram calculados utilizando o *software* Padel e a estes descritores foram adicionados outros parâmetros físico-químicos calculados pela plataforma *SwissADME*, totalizando 180 descritores iniciais e que foram normalizados, centrados na média e recalculados com variância 1. Descritores com correlação linear acima de 0.9 foram eliminados da matriz dos dados, restando nesta etapa 67 descritores.

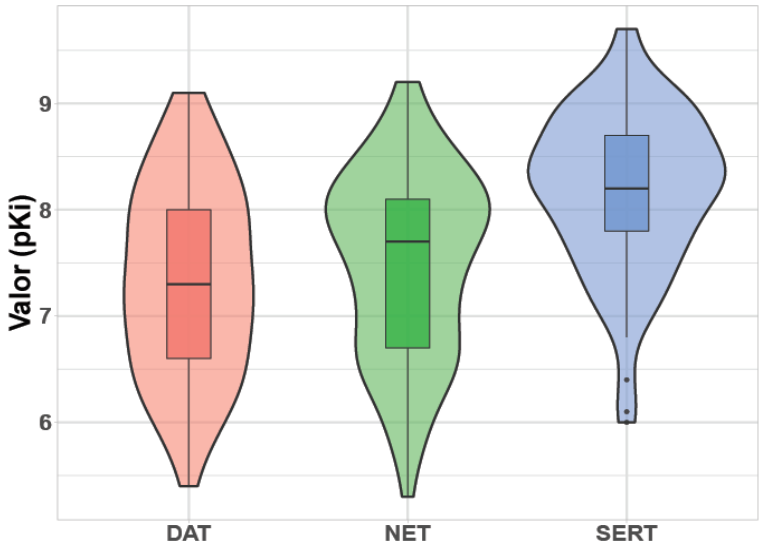
Os modelos não-lineares foram obtidos empregando os algoritmos RF e SVM, este último para efeito de comparação com os resultados obtidos com o RF, dada sua robustez para respostas não-lineares.

3. RESULTADOS

4.1 Estrutura dos dados

A distribuição dos dados relacionados a atividade biológica foi analisada, separados pelos respectivos alvos biológicos: Transportador de Dopamina (DAT), Transportador de Norepinefrina (NET) e Transportador de Serotonina (SERT). A figura 5 representa a distribuição dos valores para a inibição *in vitro* dos transportadores monoaminérgicos. O gráfico indica que os dados de inibição de dopamina são aparentemente simétricos, indicando uma distribuição próxima a normalidade. Na tabela 1, é possível observar que, para um intervalo de confiança de 90% não rejeitamos a hipótese nula de normalidade para os dados relacionados a este receptor. No entanto, com relação aos demais receptores, não podemos afirmar se a distribuição assume normalidade para um $\alpha = 10\%$.

Figura 5 - Distribuição dos valores da atividade inibitória das moléculas utilizadas. Legenda: DAT: Transportador de dopamina; NET: Transportador de norepinefrina; SERT: Transportador de serotonina; as regiões sombreadas correspondem a densidade de distribuição dos pontos e a região central é mostrado o gráfico boxplot correspondente.



Fonte: Elaborado pelo autor.

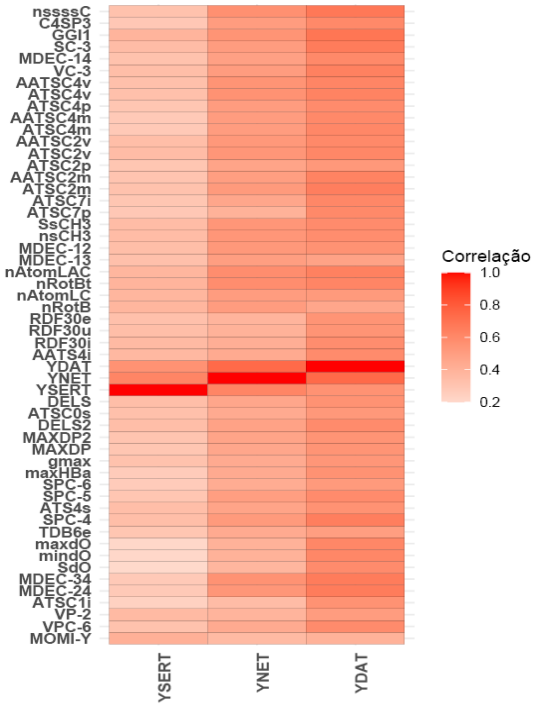
Tabela 1 - Testes para normalidade dos dados, com α = 0.1. A normalidade é averiguada para os valores relacionados à inibição de recaptura de dopamina.

Teste	Valor p - SERT	Valor p - DAT	Valor p - NET
Shapiro-Wilk	0.063	0.278	0.083
Anderson-Darling	0.096	0.325	0.023

Fonte: Dados compilados baseados nos valores de LUCAS et al., 2009; CARTER et. al 2010; LUCAS et al., 2010

Para o cálculo dos descritores 2D e 3D, as estruturas em formato ".mol" foram carregadas no *software* Padel. Utilizando o arquivo em ".csv" obtido com a matriz de descritores calculados, retirou-se aqueles que apresentavam variância próxima de 0 utilizando *software* KNIME. Com estes dados foi elaborado um *heatmap* (figura 6) dos descritores que apresentaram as maiores correlações com as três respostas biológicas individualmente. Descritores com correlação linear de Pearson acima de 0,5 com as respostas para NET e DAT, e acima de 0,4 no caso de SERT foram mantidos, totalizando 66 descritores.

Figura 6 - *Heatmap* dos descritores com as maiores correlações lineares com as 3 atividades biológicas, simultaneamente.



Legenda: quanto mais próxima do vermelho é a cor da célula correspondente ao par de descritores, mais próxima de 1 é a correlação. YDAT: Resposta biológica para o transportador de dopamina; YNET: Resposta biológica para o transportador de norepinefrina; YSERT: resposta biológica para o transportador de serotonina. Fonte: Elaborado pelo autor.

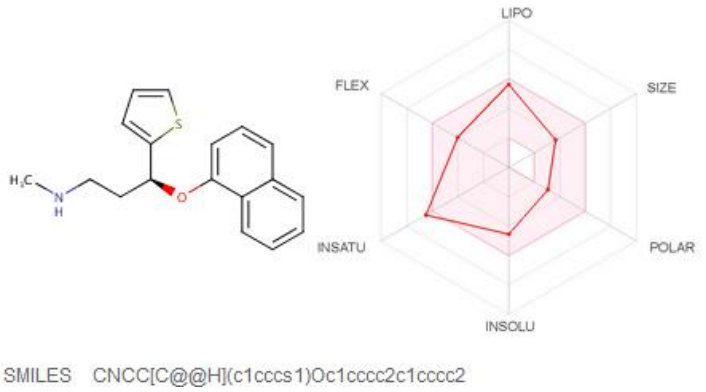
4.2 Propriedades farmacocinéticas

A plataforma *SwissADME* é munida de diversos métodos para a predição da solubilidade de um composto em sistema aquoso: ESOL (*Estimated SOLubility*), ALI e *Silicos IT*. O primeiro é um método utilizado para estimar a solubilidade aquosa de um composto derivado diretamente de sua estrutura bidimensinal. ALI é uma modificação da chamada "*General Solubility Equation*", que por sua vez é um modelo quantitativo baseado no logP e ponto de fusão de uma substância para estimar a solubilidade de compostos não ionizáveis. Dada a limitação particularmente do ponto de fusão do composto, Ali et al., propuseram a substituição do parâmetro físico-químico ponto de fusão e substituíram pelo parâmetro Área Superficial Polar Topográfica (ALI et al., 2012).

Abaixo, juntamente com a estrutura bidimensional, o gráfico de radar revela propriedades e parâmetros relevantes do ponto de vista farmacocinético.

Entre os parâmetros calculados estão propriedades físico-químicas como lipofilicidade (calculada a partir de diversas metodologias pela plataforma), solubilidade e classe de solubilidade da molécula, classe farmacocinética, potencialidade de ser um candidato a fármaco (*leadlikeness*) e alertas do ponto de vista de facilidade sintética e de química farmacêutica.

Figura 7 - Exemplo de saída dos resultados individuais das moléculas, nesse caso a Duloxetina, inserida na plataforma *SwissADME*. A área sombreada de rosa representa os valores limites ideais do respectivo parâmetro farmacocinético.



Legenda: LIPO (Lipofilicidade): XLOGP3 = entre -0.7 e 5; SIZE: entre 150 e 500g/mol; POLAR (Polarizabilidade): TPSA = entre 20 e 130A²; INSOLU (Solubilidade): logS = não maior que 6; INSATU(Saturação): fração de carbonos sp³ na molécula não menor que 0.25; FLEX(Flexibilidade): não mais que 9 ligações rotacionáveis.
Fonte: DAINA et al., 2017.

Tabela 2 - Classificação dos compostos em relação a solubilidade em diferentes métodos utilizados pela plataforma *SwissADME*

Método	Classe	N
ESOL	Moderadamente Solúvel	43
ESOL	Solúvel	34
Ali	Moderadamente Solúvel	40
Ali	Pouco solúvel	02
Ali	Solúvel	34
Silicos IT	Moderadamente Solúvel	22
Silicos IT	Pouco solúvel	54
Silicos IT	Solúvel	01

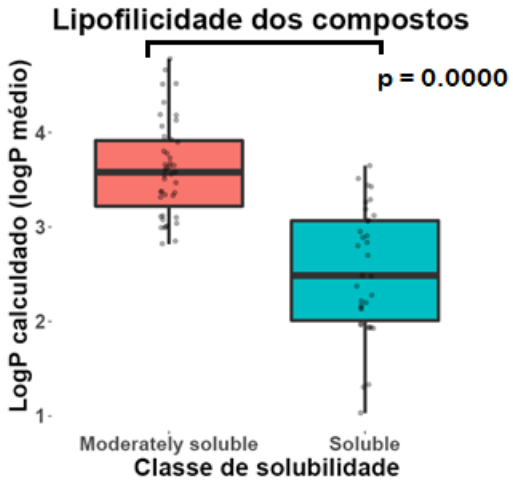
Fonte: Dados do autor obtidos pela plataforma *SwissADME* (DIANA et al., 2017)

É possível observar, na tabela, que boa parte das moléculas presentes no *dataset* foram consideradas “moderadamente solúvel” pela maioria dos métodos

empregados, com exceção do método *Silicos IT*, no qual a maioria dos compostos é classificada como “pouco solúvel”. Isso significa que é possível que boa parte das moléculas presentes no estudo apresentem boa disponibilidade oral, uma condição extremamente favorável no início desenvolvimento de um novo medicamento. Segundo o parâmetro “*leadlikeness”*, que reflete a capacidade de uma estrutura possuir características essenciais a um fármaco, 32 estruturas apresentaram nenhuma violação e 45 estruturas com apenas uma violação. Essas informações demonstram que nosso conjunto de dados é, em linhas gerais, adequado para o planejamento de novos candidatos à fármaco.

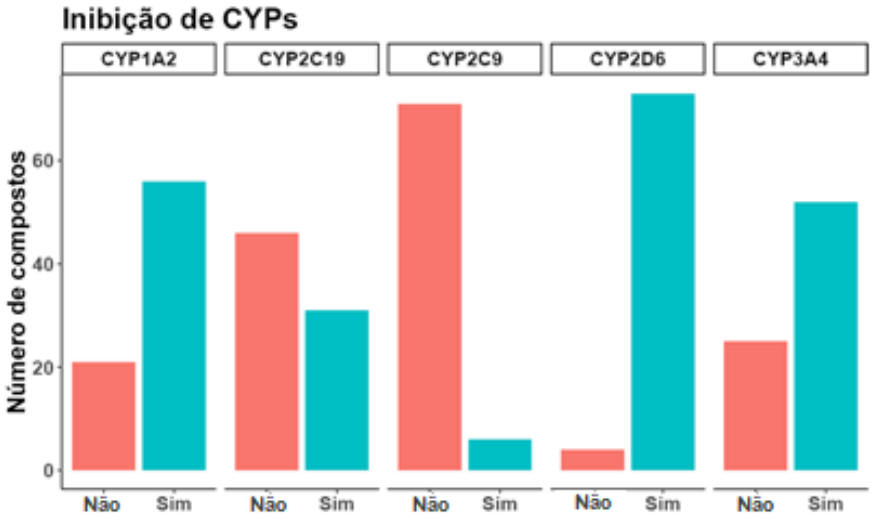
Foi investigado se moléculas que foram alocadas em diferentes classes de solubilidade possuíam logP médio significativamente diferentes. A média do grupo classificado como moderadamente solúvel é diferente do grupo classificado como solúvel num intervalo de confiança de 99%, conforme é possível observar na figura 8.

Figura 8 - Faixa dos valores de logP e classe de solubilidade pelo método ESOL



Fonte: Elaborado pelo autor.

Figura 9 - Predição da inibição de CYPs com base nas estruturas das moléculas, calculado pela plataforma SwissADME.



Fonte: Elaborado pelo autor.

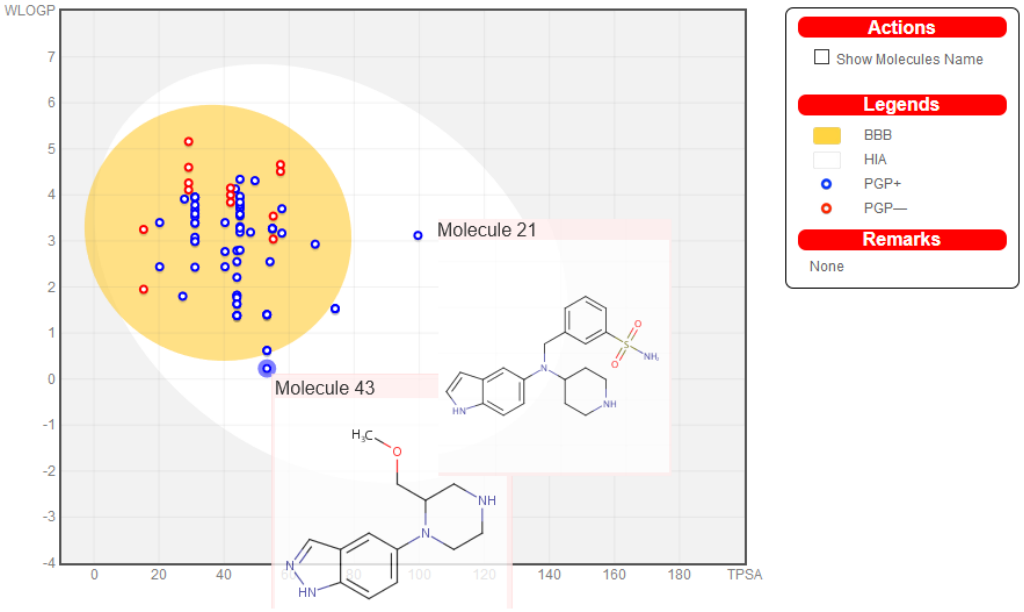
SwissADME permite a estimativa de um determinado composto ser ou não substrato da proteína de efluxo P-gp, bem como se possui capacidade de inibição de diversas isoenzimas da família CYP. O algoritmo utilizado pelo programa para esta classificação binária (substrato/não substrato) é o *Support Vector Machine*, uma técnica de aprendizado de máquina amplamente utilizada na literatura em QSAR. Essa metodologia de predição foi validada por meio da classificação de banco de moléculas com 14348 estruturas, entre elas substratos e não-substratos já conhecidos dessas enzimas (DAINA et al., 2017).

BOILED-Egg do inglês "*Brain Or IntestinaL EstimateD Permeation Predictive Model"*, é um método de predição (que também pode ser visualizado de forma direta e intuitiva) da capacidade de uma determinada molécula em ser absorvida pelo trato gastrointestinal e passar a barreira hematoencefálica. A partir de dois parâmetros físico-químicos calculados, nomeadamente lipofilicidade (logP) e superfície de área polar (PSA) de um determinado composto, é possível inferir sua capacidade de ser absorvido no trato gastrointestinal ou ainda atravessar a barreira hemato-encefálica. Na realidade, é possível identificar de uma maneira visual a região ótima dos valores de logP

e PSA na qual esses critérios são atendidos. Essas regiões, quando coloridas e combinadas de forma visual aparentam um ovo cozido (DAINA et al., 2016).

É possível averiguar na figura 10 de que maneira as estruturas presentes no estudo estão inseridas nas regiões que sugerem a possibilidade de absorção gastrointestinal e capacidade de atravessarem a barreira hematoencefálica, correspondendo respectivamente às regiões branca e amarela. Esse resultado é importante para o presente estudo, uma vez que compostos com atividade antidepressiva devem atravessar a BHM para que possam exercer sua função biológica.

Figura 10 - BOILED-Egg, representação gráfica das moléculas que potencialmente passam a barreira hemato-encefálica e/ou são absorvidas no trato gastrointestinal.



Legenda: A forma em amarelo delimita a região que estruturas possuem potencialmente capacidade de atravessar a barreira hematoencefálica. A figura branca delimita a região na qual estruturas potencialmente possuem capacidade de serem absorvidas pelo trato gastrointestinal; Pontos em vermelho dizem respeito à alta possibilidade de serem substrato da P-gp, e os pontos em azul as estruturas com alta possibilidade de não serem substrato. Fonte: DAINA et. al, 2016 (Adaptado para os dados do autor)

4.3 Método dos Mínimos Quadrados Parciais

O método dos Mínimos Quadrados Parciais (PLS), do inglês "*Partial Least Squares"* foi utilizado na construção de modelos lineares de QSAR. É baseado na redução de dimensionalidade na matriz de dados, no qual se obtém as chamadas componentes principais, satisfazendo duas condições simultâneamente: as componentes são correlacionáveis com as atividades biológicas ou variável resposta em questão; as componentes retêm o máximo da variância da matriz de descritores possível. Cada componente, que é uma combinação linear dos descritores, agora é uma dimensão e, portanto, a redução da dimensionalidade dos dados se torna possível mantendo o máximo de variância possível (HASEGAWA; FUNATSU, 2000).

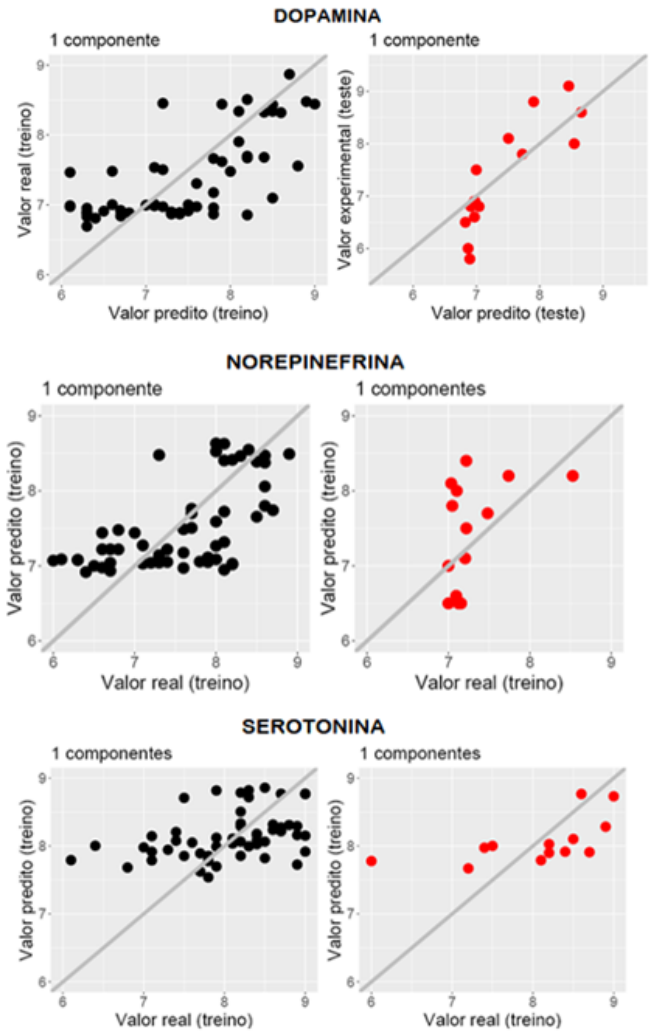
Tabela 3 - Resultados dos modelos lineares obtidos por PLS

Transportador	Grupo	# Componentes	Variância X%	R^2	*RMSE*	*MAE*
Dopamina	**Treino**	1	84.09	0,449	0,649	0,528
Dopamina	**Teste**	1	84.09	0,740	0,541	0,430
Norepinefrina	**Treino**	1	84.20	0,397	0,698	0,585
Norepinefrina	**Teste**	1	84.20	0,213	0,636	0,540
Serotonina	**Treino**	1	60.89	0,238	0,630	0,503
Serotonina	**Teste**	1	60.89	0,350	0,733	0,585

Fonte: Dados do autor.

O número de componentes foi selecionado com base na diminuição do valor de *RMSE*, ou seja, quando não se observa diferença significativa na diminuição do desvio padrão do resíduo, mantém-se o número de componentes. A variância em X diz respeito a porcentagem total de variância (leia-se informação) que é mantida em relação à matriz de descritores ou de variáveis independentes. Pode ser entendida como a quantidade de "informação" mantida, utilizando no caso do PLS a variância como parâmetro para tal medida.

Figura 11 - Valores preditos utilizando o método dos mínimos quadrados parciais.



Fonte: Elaborado pelo autor.

A tabela 3 apresenta métricas inadequadas para um modelo de QSAR, com R^2 menores que 0.5 (com excessão do grupo teste no modelo para o transportador de dopamina), assim como *RMSE* na faixa de 0.5-0.7 para todos os modelos. Esses valores observados estão abaixo do necessário para que um modelo seja considerado apropriado (KIRALJ; FERREIRA, 2009).

Tendo em vista a condição de não-linearidade da resposta, alternativas foram empregadas visando obter respostas mais fidedignas à realidade do nosso conjunto de dados. Desta forma, foram adotadas estratégias robustas a não-linearidade e para isso utilizou-se, inicialmente, o algoritmo *Random Forest* na construção de um modelo QSAR com capacidade preditiva adequada.

4.4 *Random Forest* – Modelo completo

Uma pré-seleção de descritores foi realizada utilizando um modelo inicial, contendo os 264 descritores revelando uma performance indicativa de alta capacidade de predição. Porém, do ponto de vista de interpretabilidade se torna inviável, uma vez que os descritores presentes não oferecem uma relação de causa e efeito clara. Além disso observou-se o risco a um possível *overfit* dos dados, pois o número de variáveis é muito maior que o número de amostras.

Tabela 4 - Resumo das métricas dos modelos sem o filtro de descritores, utilizando algoritmo *Random Forest*

Transportador	Grupo	#Descritores	R^2	RMSE	MAE
Dopamina	Treino	267	0,944	0,287	0,231
Dopamina	Teste	267	0,854	0,419	0,379
Norepinefrina	Treino	267	0,941	0,297	0,235
Norepinefrina	Teste	267	0,629	0,468	0,393
Serotonina	Treino	267	0,925	0,301	0,219
Serotonina	Teste	267	0,120	0,605	0,496

Fonte: Dados do autor.

Após a obtenção do modelo inicial, as variáveis foram avaliadas conforme sua ordem de importância e posteriormente selecionadas aquelas que apresentavam alto grau de pureza/importância para predição de atividade no modelo. No modelo final, considerou-se apenas as primeiras variáveis para cada transportador, em ordem de importância atribuídas pelo próprio algoritmo. Dessa forma, o modelo não-linear obtido por esta técnica foi resultado de duas iterações: a primeira para a seleção dos melhores descritores, e a segunda, já com estes filtrados. Esse último, foi considerado o mais adequado, dada a redução significativa e concomitante preservação do *RMSE* do grupo teste e pela significativa redução do número de variáveis.

Tabela 5 - Resumo das métricas dos modelos com descritores já filtrados, utilizando algoritmo *Random Forest*

Transportador	Grupo	#Descritores	R^2	RMSE	MAE
Dopamina	Treino	12	0,894	0,324	0,257
Dopamina	Teste	12	0,834	0,388	0,360
Norepinefrina	Treino	11	0,913	0,310	0,241
Norepinefrina	Teste	11	0,737	0,398	0,349
Serotonina	Treino	12	0,907	0,299	0,226
Serotonina	Teste	12	0,126	0,644	0,508

Fonte: Dados do autor.

Para uma análise mais detalhada e visual da performance do grupo treino e teste obteve-se os gráficos de valores preditos *versus* experimentais para os modelos finais. É importante destacar que os respectivos R^2 de cada grupo não representam isoladamente uma boa métrica para avaliação do modelo e, portanto, o ângulo da reta formada no gráfico da relação predito e experimental também deve ser levado em consideração.

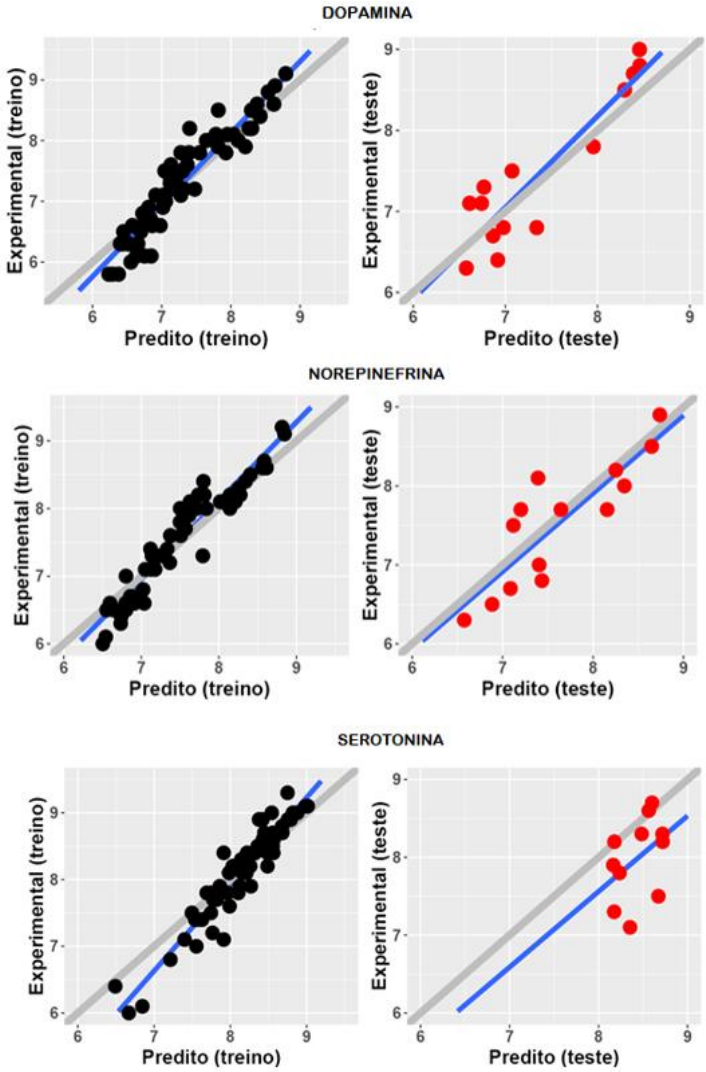
Os argumentos do algoritmo, nomeadamente o “número de árvores” e o “*mtry*” (relacionado ao número de descritores testado em cada nodo da Árvore de Decisão) foram previamente ajustados de maneira automática pelo próprio algoritmo. Os resultados estão dispostos na tabela 6.

Tabela 6 - Argumentos do método *Random Forest* utilizado para obter os modelos completos iniciais (antes da etapa de seleção dos descritores) e finais

Modelo	#Descritores	Mtry	Número de Árvores
Dopamina(pré-filtro)	267	83	1175
Dopamina(final)	12	02	116
Norepinefrina(pré-filtro)	267	19	915
Norepinefrina(final)	11	02	915
Serotonina(pré-filtro)	267	18	915
Serotonina(final)	12	02	915

Fonte: Dados do autor

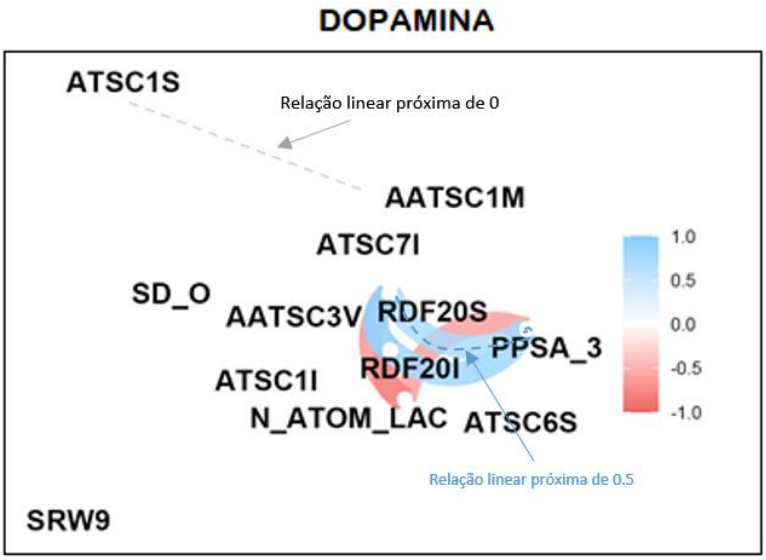
Figura 12 - Relação entre valores preditos e experimentais para a atividade inibitória de recaptura de dopamina, norepinefrina e serotonina.



Fonte: Elaborado pelo autor.

Também se mostrou necessário, para a confirmação da seleção dos descritores, a análise posterior se aqueles considerados no modelo final apresentavam multicolinearidade. Para tanto, obteve-se o gráfico de relação linear, do tipo "*network-plot*", no qual as relações lineares entre as variáveis estão dispostas de maneira visual.

Figura 13 - Relação linear entre os descritores obtidos no modelo para atividade de inibição do transportador de dopamina.



Legenda: descritores com relações lineares (positivas e negativas) estão mais próximos. Relação linear positiva é azul enquanto relação linear negativa é representada vermelha, sendo a intensidade e opacidade da cor proporcional a intensidade da relação.
Fonte: Elaborado pelo autor.

4.5 *Random Forest* - descritores interpretáveis

Um modelo inicial foi obtido, utilizando os 67 descritores com alguma interpretabilidade química, como: número de carbonos presentes, ligações duplas, ligações rotacionáveis, estado eletrônico de heteroátomos e entre outros. Destes considerados, foram filtrados os mais relevantes para a predição da atividade, relevância atribuída conforme método de impureza do próprio algoritmo para, finalmente, se obter o modelo final.

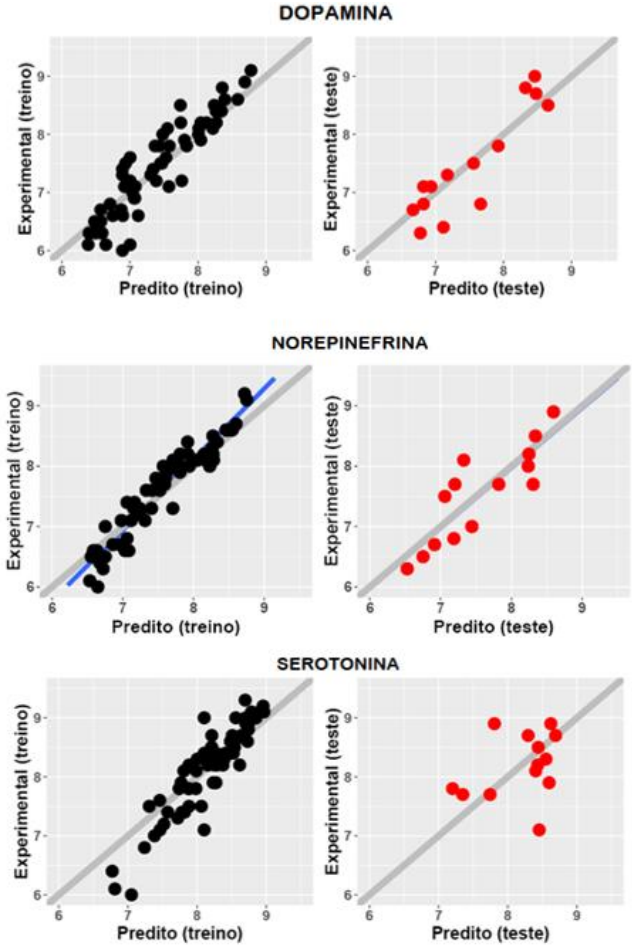
A tabela 7 dispõe as métricas utilizando todos os descritores iniciais para a construção de um modelo inicial. Este, por sua vez, foi utilizado para filtrar os principais descritores em ordem de importância atribuídas por RF, e a partir de então reduzir a dimensionalidade da matriz de descritores.

Tabela 7 - Resumo das métricas obtidas no modelo inicial

Transportador	Grupo	#Descritores	R^2	RMSE	MAE
Dopamina	Treino	67	0,898	0,335	0,264
Dopamina	Teste	67	0,832	0,436	0,374
Norepinefrina	Treino	67	0.935	0.293	0.230
Norepinefrina	Teste	67	0.712	0.406	0.350
Serotonina	Treino	67	0,89	0,339	0,261
Serotonina	Teste	67	0,075	0,593	0,44

Fonte: Dados do autor

Figura 14 - Relação entre valores preditos e experimentais para a atividade inibitória de recaptura de dopamina.



Legenda**:** À esquerda está a relação entre os valores experimentais e preditos referentes às estruturas utilizadas na construção do modelo, ou grupo treino. À direita está a relação entre os valores experimentais e preditos referentes às estruturas não utilizadas na construção do modelo, ou grupo teste. Fonte: elaborado pelo autor.

Obteve-se finalmente, um modelo utilizando os descritores filtrados anteriormente e as respectivas métricas estão dispostas na tabela 8.

Tabela 8 - Resumo das métricas obtidas no modelo final, com os descritores filtrados

Transportador	Grupo	#Descritores	R^2	RMSE	MAE
DOPAMINA	TREINO	10	0,863	0,362	0,277
DOPAMINA	TESTE	10	0,817	0,398	0,304
NOREPINEFRINA	TREINO	15	0,922	0,301	0,239
NOREPINEFRINA	TESTE	15	0,737	0,388	0,337
SEROTONINA	TREINO	9	0,847	0,365	0,285
SEROTONINA	TESTE	9	0,104	0,586	0,437

Fonte: Dados do autor

4.5.1 *Support Vector Machine* – com todos os descritores

Com o propóstio de explorar a técnica, bem como comparar diferentes métodos para abordagem do problema, foram elaborados modelos a partir do método de SVM do tipo radial, com 267 descritores iniciais, os mesmos utilizados no modelo RF completo. As métricas de performance podem ser observadas na tabela 9.

Tabela 9 - Métricas obtidas utilizando Support Vector Machine

Modelo	Grupo	R^2	RMSE	MAE
Dopamina	Treino	0,811	0,395	0,224
Dopamina	Teste	0,799	0,418	0,355
Norepinefrina	Treino	0,784	0,458	0,355
Norepinefrina	Teste	0,601	0,487	0,406
Serotonina	Teste	0,024	0,613	0,503
Serotonina	Treino	0,845	0,486	0,388

Fonte: Dados do autor

4. DISCUSSÃO

5.1 Escolha dos softwares

KNIME, do inglês “*Konstanz Information Miner*” é um *software* que possui ambiente modular e visual, utilizado na elaboração de fluxos de processamento, manipulação e visualização de dados. Este *software* integra diversos algoritmos e ferramentas utilizadas para problemas de regressão/predição ou classificação de dados. Opera a partir de “nós” ou ”nodos” que representam etapas unitárias e modulares, as quais possuem funções específicas na manipulação e processamento dos dados e que podem ser configurados conforme o objetivo do usuário. Um nó é interligado a diversos outros nós e é possível, dessa maneira, automatizar longos processos. Uma de suas principais vantagens, além de se tratar de um software *open-source* (i.e., uso gratuito) é a sua característica visual e intuitiva, sem necessitar de conhecimentos aprofundados e específicos de programação. Utilizando o KNIME é possível realizar, por exemplo, de maneira muito simples e automatizada, a conversão entre os diversos formatos de arquivos para representação estrutural de moléculas como SMILES ou MOLFile. (BERTHOLD, 2009). Este *software* foi utilizado para a transformação entre os diversos formatos das estruturas (SMILES, MOLFile) bem como utilizado no filtro inicial dos descritores com variância próxima de (0) zero.

Para o cálculo dos descritores foi utilizado o *software* Padel, que por sua vez, também é um gratuito e oferece uma interface em Java bastante simples e amigável para o usuário.

A linguagem de programação R possui grande relevância no contexto de quimiometria, devido à sua ampla gama de funcionalidades e ferramentas presentes nos diversos pacotes presentes. Estes pacotes ou bibliotecas, do inglês “*library*” são geralmente criados pela própria comunidade usuária da linguagem, bem como também é gratuito e totalmente acessível em diferentes sistemas operacionais (WEHRENS, 2010). Todos os modelos para predição e gráficos utilizados neste trabalho foram obtidos utilizando a linguagem R.

5.2 Estrutura e análise inicial dos dados

É possível observar na figura 5 que a distribuição dos valores está de acordo com as boas práticas de desenvolvimento de modelos QSAR, abrangendo valores na faixa de, pelo menos, 3 unidades logarítmicas. O *boxplot* dos valores de inibição também nos permite chegar à mesma conclusão e ele nos dá ainda mais uma informação relevante, de que é possível que existam *outliers* com relação aos valores obtidos de pKi e que são um alerta na condução do estudo.

Observa-se no *heatmap* (figura 6) que as maiores correlações lineares estão presentes para a resposta biológica ligada ao transportador de dopamina uma vez que visualmente as cores estão mais próximas de tons de vermelho nas células. Essa informação é de grande relevância na construção de um modelo assumindo-se relação linear entre os descritores e a variável dependente, visto que é possível visualizar no próprio *heatmap* que as relações lineares entre os descritores e as respostas biológicas não são tão intensas. É importante destacar que correlação, nesse caso estritamente linear, não significa necessariamente causalidade, porém nos dá uma ideia de como os descritores presentes estão relacionados entre si. Dadas essas informações iniciais, é esperado um modelo linear mais robusto utilizando a resposta dopaminérgica, o que realmente foi confirmado ao longo do desenvolvimento do estudo, dadas as maiores métricas utilizadas na validação dos modelos obtidos – R^2, RMSE e MAE (tabela – para os grupos teste e treino no modelo linear para este transportador.

Dada a condição de não-linearidade, que não é incomum na literatura, empregou-se, posteriormente, técnicas adequadas com o objetivo de construir um modelo robusto a relações não-lineares entre a variável dependente (atividade antidepressiva *in vitro*) e variáveis independentes/matriz de descritores (MICHIELAN; MORO, 2010).

5.3 Análise dos dados farmacocinéticos

Todas as moléculas presentes nesse estudo apresentaram alta absorção gastrointestinal, classificados como “*High*” no parâmetro “*GI Absorption”*, além

de alta permeação da barreira hematoencefálica, ou seja, 73 das 77 moléculas foram classificadas como "*BBB permeant*". Tal parâmetro farmacocinético é de grande valia para um potencial candidato a fármaco com atividade antidepressiva, visto que seu mecanismo de ação se dá no sistema nervoso central. Por se tratar de uma classe de compostos com atividade antidepressiva *in vitro*, as possíveis capacidades de absorção no trato gastrointestinal bem como permeabilidade da barreira hematoencefálica mostram-se inicialmente promissoras.

Após o cálculo e processamento de todas as moléculas, é possível visualizar de forma interativa os resultados que dizem respeito à predição da capacidade de atravessar a barreira hematoencefálica e facilidade de absorção gastrointestinal, contemplados neste trabalho (DAINA et al., 2017).

Alguns compostos presentes no estudo foram classificados pelo método ESOL da plataforma *SwissADME* como moderadamente solúvel com valores de logP calculados que entram na faixa de 3 a 5, e o restante classificados como solúveis, com valores que vão de 1 a aproximadamente 3,5 ($p = 1^{-10}$ para o teste-t de Welch com intervalo de confiança de 95% e hipótese nula da diferença das médias ser zero).

Em relação à possível inibição de enzimas da família CYP, foi observado que a maior parte dos compostos possui potencial perfil inibitório para *CYP1A2*, *CYP2D6* e *CYP3A4*, o que representaria um risco de interação entre diversos medicamentos que são extensivamente metabolizados por estas vias. Não obstante, o conhecimento sobre os compostos serem substratos ou não substratos da glicoproteína P é fundamental para entender se há a possibilidade de efluxo ativo da parede gastrointestinal para o lúmen ou, também, pelo ativo do sistema nervoso central, comprometendo diretamente o acesso do fármaco ao alvo molecular e seu mecanismo de ação (VAN WATERSCHOOT et al., 2011).

5.4 Análise do Método Mínimos Quadrados Parciais

É possível observar pelos valores de R^2, RMSE e MAE dos modelos obtidos que todos performaram de maneira insatisfatória, assumindo-se relações

lineares entre a variável biológica e os descritores. Tanto para o grupo treino quanto para o grupo teste os modelos não são preditivos e, portanto, não é possível inferir nenhuma informação relevante do ponto de vista químico e/ou estrutural. Pode se dizer que, por se tratar de $R^2 < 0.5$ e os altos valores de RMSE, correlações são espúrias entre os descritores selecionados e a atividade predita .

Como exemplo da aplicação de QSAR na predição de toxicidade VOTANO (2004), dada a grande variedade de mecanismos e interações, a resposta biológica nem sempre será próxima da linear. Com a finalidade de estabelecer modelos não-lineares, o autor empregou técnicas alternativas como Redes Neurais Artificiais e k-Vizinho-Mais-Próximo, sendo pioneiro no emprego de técnicas não-lineares na predição de toxicidade. Neste trabalho utilizou-se a metodologia *Random Forest* e SVM para abordar o problema.

5.5 Discussão de Resultados Modelo Completo

No modelo construído em relação à dopamina (figura 12) observa-se que a linha de tendência azul está levemente deslocada em relação a linha cinza (a qual representa se obtivéssemos um modelo teórico perfeito) no grupo treino, porém significativamente semelhante no grupo teste. Um ângulo muito agudo, por exemplo, pode representar uma superestimação de valores mais altos e subestimação de valores experimentais mais baixos e mesmo assim estar ligado a um R^2 alto. Uma métrica informativa e que contorna esse problema é o CCC, do inglês "*Concordance Correlation Coeficient"* (CHIRICO, GRAMATICA, 2011).

Podemos observar os três descritores mais importantes em cada modelo e seu respectivo significado físico-químico na tabela abaixo (tabela 10) para compreender quais descritores influenciam na atividade, bem como se há uma convergência de descritores em comum e simultaneamente relevantes na predição de atividade em diferentes transportadores.

Tabela 10 - Resumo das métricas obtidas no modelo final, com os descritores filtrados. Descritores que aparecem em mais de um modelo estão assinalados

Transportador	Descritor	Significado
Dopamina	**RDF20I**	Função de distribuição radial – 020 / ponderado pelo potencial relativo de primeira ionização
Dopamina	**RDF20S**	Função de distribuição radial – 020/ ponderada pelo I-Estado relativo
Dopamina	ATSC6S	Autocorrelação de Broto-Moreau centrada – lag 6/ ponderada pelo I-Estado
Norepinefrina	**RDF20I**	Função de distribuição radial – 020/ ponderada pelo primeiro potencial de ionização relativo
Norepinefrina	**RDF20S**	Função de distribuição radial – 020/ ponderada pelo I-Estado relative.
Norepinefrina	VE3_DT	Coeficiente logarítmico da soma do último *eigenvector* da matriz de desvio
Serotonina	RDF40E	Função de distribuição radial - 040/ ponderado pela Eletronegatividade relativa de Sanderson
Serotonina	RDF40V	Função de distribuição radial - 040/ ponderada pelo volume relativo de van der Waals
Serotonina	**RDF20S**	Função de distribuição radial - 020 / ponderada pelo I-Estado relativo

Fonte: TODESCHINI; CONSONNI, (2009)

As métricas obtidas diferem de maneira significativa entre o grupo teste e treino, e é possível afirmar que o modelo de serotonina não obteve performance esperada e trata-se de um *overfitting*. À medida que se aumenta a complexidade do modelo e dos algoritmos utilizados é possível que ocorra um fenômeno conhecido como *overfit* dos dados. Isso acontece pois o modelo essencialmente se adequa aos erros ou ruídos associados aos dados o que resulta, em última instância, numa baixa performance na predição de dados externos que não fazem parte da série treino (JAMES et. al, 2017).

Podemos destacar dois descritores importantes que apresentam importância em ambos os modelos, sendo tanto para a predição de atividade frente aos receptores de dopamina e norepinefrina. Destacam-se os descritores RDF20S presentes em todos os modelos obtidos e o descritor RDF20I, presente nos modelos obtidos para Dopamina e Serotonina, apesar deste último não possuir valores adequados na predição de dados externos. Formalmente, a função de distribuição radial pode ser interpretada como a distribuição de probabilidade em encontrar um átomo num volume esférico de raio R (TODESCHINI; CONSONNI, 2009). Do ponto de vista químico sua interpretação

se torna bem inacessível e portanto, na tentativa de gerar modelos com descritores mais interpretáveis, prosseguiu-se com a eliminação de descritores com significados químicos mais objetivos.

Em relação aos resultados apresentados pelo modelo para o transportador de serotonina há claramente um *overfitting* dos dados treino e não há aplicabilidade nos dados externos, tendo em vista a diferença na performance externa (KIRALJ; FERREIRA, 2009).

No que se refere à complexidade das informações contidas nos descritores, a interpretação do modelo se torna demasiadamente complicada. Por exemplo, os descritores presentes no modelo para dopamina como ATSC6S, AATSC3V, ATSC7I, ATSC1l são bidimensionais e relacionados a autocorrelação, baseados em teoria dos grafos, por sua vez resultantes da estrutura das moléculas presentes, o que torna a interpretação do modelo muito difícil do ponto de vista químico ou mecanístico.

Apesar de atingir os objetivos de capacidade de predição dos dados externos para os transportadores de dopamina e norepinefrina, em vista da complexidade da intepretação adequada do modelo, optou-se pela construção de um segundo modelo, lançando mão da mesma técnica, porém utilizando apenas descritores que sejam mais simples e interpretáveis, mantendo-se ainda boa qualidade preditiva.

Com o objetivo de comparar diferentes métodos, o modelo SVM obtido com os 267 descritores iniciais, não foi superior em performance quando comparado ao modelo obtido por RF.

5.6 Discussão do Modelo RF com descritores interpretáveis

Foi possível averiguar que as métricas de performance (R^2, RMSE) não apresentaram mudanças drásticas entre o modelo com 67 descritores e o modelo com apenas os mais importantes do modelo anterior e, portanto, pelo princípio da parcimônia ou “navalha de Ockham”, o qual estabelece que um número menor de descritores que fornece uma melhor interpretabilidade e capacidade de generalização (HOFFMAN, et. al, 1996), optamos pelo modelo com um número menor de descritores.

Tabela 11 - Resumo das métricas obtidas no modelo final, com os descritores filtrados

Transportador	Descritor	Significado
Dopamina	N_ATOM_LAC	Número de átomos na maior cadeia alifática
Dopamina	GMIN	E-Estado mínimo
Dopamina	**MDEC_14**	Distância molecular entre todos os carbonos primários e quaternários
Dopamina	**SSSS_N**	Soma do E-Estado para: >N-
Dopamina	SH_BA	Soma de estados-E para aceptores de ligação de hidrogênio
Norepinefrina	**SSSS_N**	Soma do E-Estado para: >N-
Norepinefrina	N_ROT_BT	Número de ligações rotacionáveis, incluindo ligações terminais
Norepinefrina	X_LOG_P	Lipofilicidade Calculada
Norepinefrina	**MDEC_14**	Distância molecular entre carbonos primários e quaternários
Norepinefrina	MAX_H_BA	E-Estado máximo para receptores de ligação de hidrogênio
Serotonina	ETA_ETA	Composite index Eta
Serotonina	N_ATOM_P	Número de átomos no maior Sistema pi aromático
Serotonina	ALI_SOLUBILITY	Solubilidade calculada pelo método ALI

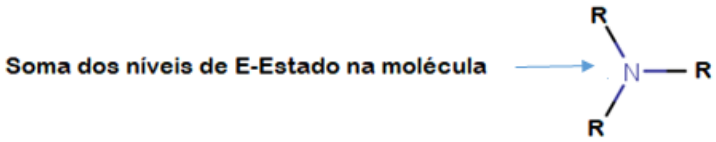
Fonte: TODESCHINI; CONSONNI, (2009)

Na tabela acima destaca-se o descritor “N_ATOM_LAC” que se refere ao número de átomos cadeia alifática principal, indicando possivelmente um tamanho ótimo na cadeia para a interação com o transportador de dopamina. “GMIN” descreve o E-Estado mínimo da molécula, e é uma medida relacionada ao átomo mais eletrofílico na estrutura. Esses dois descritores físico-químicos possuem alta relevância para a predição de atividade biológica da molécula. O descritor “SH_BA” que está relacionado a fortes aceptores da ligação de hidrogênio presentes na estrutura e também possui importância na predição da atividade biológica.

Para o modelo obtido para atividade no transportador de norepinefrina, destaca-se ainda o descritor “N_ROT_BT” que diz respeito ao número de ligações rotacionáveis presentes na estrutura incluindo ligações terminais, e o “XLOGP” que é uma medida de lipofilicidade calculada.

"MDEC_14" e "SSSS_N", que aparecem em ambos os modelos para dopamina e norepinefrina, estão relacionados à distância entre os carbonos primários e quaternários e o estado eletrônico dos átomos de nitrogênio presentes na estrutura, respectivamente. Para efeito de ilustração do significado, o "E-Estado" calculado para um átomo de carbono aromático é menor quando este está adjacente a um carbono substituído com um grupamento hidroxila e ainda menor quando o grupo hidroxila está diretamente ligado ao átomo de carbono, ou seja, o "E-Estado" de um átomo depende diretamente de grupos eletrofílicos próximos (VOTANO, 2004).

Figura 15 - Representação do descritor "ssssN"



Fonte: VOTANO, 2004 (Adaptado)

Não é possível realizar uma intepretação para o modelo de serotonina uma vez que suas métricas de performance, tanto para o grupo treino quanto para o grupo teste foram insatisfatórias, e, portanto, não é um modelo adequadamente validado.

5. CONCLUSÃO

Pela técnica utilizada, observou-se que modelos que empregavam *Machine Learning* em sua composição, relacionando a estrutura química e atividade biológica de forma não-linear, obtiveram performances muito superiores a técnicas lineares. No entanto, não foi possível construir um modelo adequado que relacionasse as características físico-químicas deste grupo de moléculas e atividade biológica para o transportador de serotonina. É possível que, diante da técnica empregada, não houve a captura dos mecanismos relevantes para a atividade biológica neste transportador e metodologias mais robustas de QSAR, a citar *CoMFA/CoMSIA* podem ser alternativas adequadas nesse caso, assim como a necessidade de mais estruturas químicas presentes na elaboração do modelo.

As moléculas presentes no estudo apresentaram parâmetros farmacocinéticos preditos que se mostram adequados no desenvolvimento de um novo candidato a fármaco.

Para os diferentes algoritmos e técnicas usadas, RF se mostrou adequado mesmo quando comparado a outra técnica, neste caso SVM, seja modelagem ou na predição dos dados. Os modelos RF não-lineares para dopamina e norepinefrina concordam do ponto de vista de descritores relevantes para a predição de atividade biológica, tanto no modelo utilizando a matriz completa de descritores quanto no simplificado, com destaque a descritores que aparecem em ambos os modelos:

- Função de Distribuição Radial “20I” e “20S” nos modelos com todos os descritores calculados, sugerindo uma grande importância da estrutura tridimensional na atividade biológica;
- Estado eletrônico do(s) nitrogênio(s) presente(s) na molécula;
- Distância entre carbonos primários e quaternários (quando presente) apontando para um tamanho ótimo da estrutura nos modelos com descritores facilmente interpretáveis;

Do ponto de vista de capacidade de predição, foi possível atingir os objetivos propostos na construção de modelos preditivos e com métricas de performance adequadas quanto à atividade inibitória dos transportadores de

dopamina e norepinefrina, com destaque aos modelos referentes ao primeiro e com descritores convergentes, ou seja, que aparecem em ambos os modelos. O modelo não-linear construído a partir descritores interpretáveis para o transportador de dopamina performou tão bem quanto o modelo com todos os descritores calculados e, portanto, é preferível devido a sua maior simplicidade e redução de demanda computacional.

6. REFERÊNCIAS

ALI, J. et al. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. Journal of Chemical Information and Modeling, v. 52, n. 2, p. 420–428, 13 jan. 2012. Disponível em: <http://dx.doi.org/10.1021/ci200387c>.

ALVES, V. et al. QUIMIOINFORMÁTICA: UMA INTRODUÇÃO. Química Nova, 2017. Disponível em: <http://dx.doi.org/10.21577/0100-4042.20170145>.

BERTHOLD, M. R. et al. KNIME - the Konstanz Information Miner. ACM SIGKDD Explorations Newsletter, v. 11, n. 1, p. 26–31, 16 nov. 2009. Disponível em: <http://dx.doi.org/10.1145/1656274.1656280>.

BREIMAN, L. Machine Learning, v. 45, n. 1, p. 5–32, 2001. Disponível em: <http://dx.doi.org/10.1023/A:1010933404324>.

BURBIDGE, R. et al. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. Computers & Chemistry, v. 26, n. 1, p. 5–14, dez. 2001. Disponível em: <http://dx.doi.org/10.1016/s0097-8485(01)00094-8>.

CARTER, D. S. et al. 2-Substituted N-Aryl Piperazines as Novel Triple Reuptake Inhibitors for the Treatment of Depression. Bioorganic & Medicinal Chemistry Letters, v. 20, n. 13, p. 3941–3945, 2010.

CHIRICO, N.; GRAMATICA, P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. Journal of Chemical Information and Modeling, v. 51, n. 9, p. 2320–2335, 12 ago. 2011. Disponível em: <http://dx.doi.org/10.1021/ci200211n>.

DAINA, A.; MICHIELIN, O.; ZOETE, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. Scientific Reports, v. 7, n. 1, 3 mar. 2017. Disponível em: <http://dx.doi.org/10.1038/srep42717>.

DAINA, A.; ZOETE, V. A BOILED-Egg To Predict Gastrointestinal Absorption and Brain Penetration of Small Molecules. ChemMedChem, v. 11, n. 11, p. 1117–1121, 24 maio 2016. Disponível em: <http://dx.doi.org/10.1002/cmdc.201600182>.

DAMALE, M. et al. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. Mini-Reviews in Medicinal Chemistry, v. 14, n. 1, p. 35–55, 31

jan. 2014. Disponível em: <http://dx.doi.org/10.2174/13895575113136660104>.

FERRARI, A. J. et al. Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. PLoS Medicine, v. 10, n. 11, p. e1001547, 5 nov. 2013. Disponível em: <http://dx.doi.org/10.1371/journal.pmed.1001547>.

FERREIRA, M. M. C. Multivariate QSAR. Journal of the Brazilian Chemical Society, v. 13, n. 6, nov. 2002. Disponível em: <http://dx.doi.org/10.1590/S0103-50532002000600004>.

FISKE, A.; WETHERELL, J. L.; GATZ, M. Depression in Older Adults. Annual Review of Clinical Psychology, [s. l.], v. 5, n. 1, p. 363–389, 2009. Disponível em: <http://dx.doi.org/10.1146/annurev.clinpsy.032408.153621>

GRAMATICA, P.; SANGION, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. Journal of Chemical Information and Modeling, v. 56, n. 6, p. 1127–1131, 3 jun. 2016. Disponível em: <http://dx.doi.org/10.1021/acs.jcim.6b00088>.

HASEGAWA, K.; FUNATSU, K. Partial Least Squares Modeling and Genetic Algorithm Optimization in Quantitative Structure-Activity Relationships. SAR and QSAR in Environmental Research, v. 11, n. 3–4, p. 189–209, ago. 2000. Disponível em: <http://dx.doi.org/10.1080/10629360008033231>.

HOFFMANN, ROALD & MINKIN, VLADIMIR & CARPENTER, BARRY. (1996). Ockham's Razor and Chemistry. Bulletin de la Societe Chimique de France. 133.

LANE, R. M. Antidepressant Drug Development: Focus on Triple Monoamine Reuptake Inhibition. Journal of Psychopharmacology, v. 29, n. 5, p. 526–544, 14 out. 2014. Disponível em: <http://dx.doi.org/10.1177/0269881114553252>.

LOOMER, H P et al. “A clinical and pharmacodynamic evaluation of iproniazid as a psychic energizer.” *Psychiatric research reports* vol. 8 (1957): 129-41.

LUCAS, M. C. et al. Design, Synthesis, and Biological Evaluation of New Monoamine Reuptake Inhibitors with Potential Therapeutic Utility in Depression and Pain. Bioorganic & Medicinal Chemistry Letters, v. 20, n. 18, p. 5559–5566, 2010.

LUCAS, M. C. et al. Novel, Achiral Aminoheterocycles as Selective Monoamine Reuptake Inhibitors. Bioorganic & Medicinal Chemistry Letters, v. 19, n. 16, p. 4630–4633, 2009.

MARKS, D. M.; PAE, C.-U.; PATKAR, A. A. Triple Reuptake Inhibitors: A Premise and Promise. Psychiatry Investigation, v. 5, n. 3, p. 142, 2008. Disponível em: <http://dx.doi.org/10.4306/pi.2008.5.3.142>.

MarvinSketch, versão 20.16.0, 2020 ChemAxon disponível em: <http://www.chemaxon.com>. Data de acesso: 08 agosto 2020.

MITCHELL, J. B. O. Machine Learning Methods in Chemoinformatics. Wiley Interdisciplinary Reviews: Computational Molecular Science, v. 4, n. 5, p. 468–481, 24 fev. 2014. Disponível em: <http://dx.doi.org/10.1002/wcms.1183>.

NEVES, B. J. et al. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Frontiers in Pharmacology, v. 9, 13 nov. 2018. Disponível em: <http://dx.doi.org/10.3389/fphar.2018.01275>.

NEVES, B. J. et al. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Frontiers in Pharmacology, v. 9, 13 nov. 2018. Disponível em: <http://dx.doi.org/10.3389/fphar.2018.01275>.

PEREZ-CABALLERO, L. et al. Monoaminergic system and depression. Cell and Tissue Research, v. 377, n. 1, p. 107–113, 10 jan. 2019. DOI 10.1007/s00441-018-2978-8. Disponível em: http://dx.doi.org/10.1007/s00441-018-2978-8

PEREZ-CABALLERO, L. et al. Monoaminergic System and Depression. Cell and Tissue Research, v. 377, n. 1, p. 107–113, 10 jan. 2019. Disponível em: <http://dx.doi.org/10.1007/s00441-018-2978-8>.

PICCINELLI, M.; WILKINSON, G. Gender Differences in Depression. British Journal of Psychiatry, v. 177, n. 6, p. 486–492, dez. 2000. Disponível em: <http://dx.doi.org/10.1192/bjp.177.6.486>.

SCHILDKRAUT, J. J. THE CATECHOLAMINE HYPOTHESIS OF AFFECTIVE DISORDERS: A REVIEW OF SUPPORTING EVIDENCE. American Journal of Psychiatry, v. 122, n. 5, p. 509–522, nov. 1965. Disponível em: <http://dx.doi.org/10.1176/ajp.122.5.509>.

SHARMA, H.; SANTRA, S.; DUTTA, A. Triple Reuptake Inhibitors as Potential next-Generation Antidepressants: A New Hope? Future Medicinal Chemistry, v. 7, n. 17, p. 2385–2406, nov. 2015. Disponível em: <http://dx.doi.org/10.4155/fmc.15.134>.

SVETNIK, V. et al. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In: Multiple Classifier Systems. [s.l.] Springer Berlin Heidelberg, 2004. p. 334–343

.

SVETNIK, V. et al. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In: Multiple Classifier Systems. [s.l.] Springer Berlin Heidelberg, 2004. p. 334–343.

SVETNIK, V. et al. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences, v. 43, n. 6, p. 1947–1958, nov. 2003. Disponível em: <http://dx.doi.org/10.1021/ci034160g>.

TODESCHINI, R; CONSONI, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH)

TROPSHA, A.; GOLBRAIKH, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. Current Pharmaceutical Design, v. 13, n. 34, p. 3494–3504, 1 dez. 2007. Disponível em: <http://dx.doi.org/10.2174/138161207782794257>.

VAN WATERSCHOOT, R. A. B.; SCHINKEL, A. H. A Critical Analysis of the Interplay between Cytochrome P450 3A and P-Glycoprotein: Recent Insights from Knockout and Transgenic Mice. Pharmacological Reviews, v. 63, n. 2, p. 390–410, 13 abr. 2011. Disponível em: <http://dx.doi.org/10.1124/pr.110.002584>

VOTANO, J. R. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. Mutagenesis, v. 19, n. 5, p. 365–377, 1 set. 2004. Disponível em: <http://dx.doi.org/10.1093/mutage/geh043>.

WEHRENS, R. Data. In: Chemometrics with R., Springer Berlin Heidelberg, 2010. p. 7–12.

WORLD HEALTH ORGANIZATION (2016) Out of the shadows: making mentalhealth a global development priority. World Health Organization,Washington

YAP, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. Journal of Computational Chemistry, v. 32, n. 7, p. 1466–1474, 17 dez. 2010. Disponível em: <http://dx.doi.org/10.1002/jcc.21707>.

7. ANEXOS

Anexo I. Tabela com atividade biológica in vitro e SMILE da estrutura correspondente

SMILE correspondente	pKi SERT	pKi NET	pKi DAT
CNCC[C@H](Oc1cccc2ccccc12)c3cccs3	9,30	8,10	6,60
CNCCC(c1ccccc1)c2ccc3[nH]ccc3c2	7,50	8,00	6,70
CNCCN(c1ccccc1)c2ccc3[nH]ccc3c2	6,40	6,70	6,30
C1CC(CCN1)N(c2ccccc2)c3ccc4[nH]ccc4c3	6,00	5,90	5,80
C(N(C1CCNCC1)c2ccc3[nH]ccc3c2)c4ccccc4	9,10	8,10	7,30
C(N(C1CCCNC1)c2ccc3[nH]ccc3c2)c4ccccc4	8,20	7,20	7,40
C(N(C1CCNC1)c2ccc3[nH]ccc3c2)c4ccccc4	7,70	8,00	7,50
C(N(C1CCNCC1)c2ccc3[nH]ncc3c2)c4ccccc4	8,90	7,80	7,30
C(N(C1CCNCC1)c2ccc3sccc3c2)c4ccccc4	7,90	6,70	6,70
Cn1ccc2cc(ccc12)N(Cc3ccccc3)C4CCNCC4	8,40	7,60	7,10
CC(N(C1CCNCC1)c2ccc3[nH]ccc3c2)c4ccccc4	7,40	5,80	5,80
O=C(N(C1CCNCC1)c2ccc3[nH]ccc3c2)c4ccccc4	7,60	6,10	5,40
C(Cc1ccccc1)N(C2CCNCC2)c3ccc4[nH]ccc4c3	8,20	7,30	6,80
C(C1CCOCC1)N(C2CCNCC2)c3ccc4[nH]ccc4c3	7,40	7,40	6,30
Fc1ccccc1CN(C2CCNCC2)c3ccc4[nH]ccc4c3	8,40	7,90	6,90
Fc1cc(CN(C2CCNCC2)c3ccc4[nH]ccc4c3)ccc1	8,90	7,90	7,10
Fc1ccc(CN(C2CCNCC2)c3ccc4[nH]ccc4c3)cc1	8,70	6,60	6,60
N#Cc1ccccc1CN(C2CCNCC2)c3ccc4[nH]ccc4c3	8,40	8,20	6,30
N#Cc1cc(CN(C2CCNCC2)c3ccc4[nH]ccc4c3)ccc1	9,60	8,00	7,50
N#Cc1ccc(CN(C2CCNCC2)c3ccc4[nH]ccc4c3)cc1	8,60	6,00	6,10
NS(=O)(=O)c1cc(CN(C2CCNCC2)c3ccc4[nH]ccc4c3)ccc1	8,20	7,10	7,60
COc1cc(CN(C2CCNCC2)c3ccc4[nH]ccc4c3)ccc1	8,80	8,40	7,10
C(C1CNCCN1c2ccc3[nH]ccc3c2)c4ccccc4	7,90	8,10	8,20
C(C1CNCCN1c2ccccc2)c3ccccc3	7,80	6,40	6,30
CNc1ccc(cc1)N2CCNCC2Cc3ccccc3	7,10	6,30	6,40
Clc1ccc(cc1Cl)N2CCNCC2Cc3ccccc3	7,80	7,80	7,10
Cn1ccc2cc(ccc12)N3CCNCC3Cc4ccccc4	8,10	7,30	6,90
NC(=O)c1c[nH]c2ccc(cc12)N3CCNCC3Cc4ccccc4	7,00	5,30	6,80
NC(=O)c1cc2cc(ccc2[nH]1)N3CCNCC3Cc4ccccc4	8,10	6,30	8,50
COc1c2[nH]ccc2cc(c1)N3CCNCC3Cc4ccccc4	7,30	7,30	5,80
Clc1c2[nH]ccc2cc(c1)N3CCNCC3Cc4ccccc4	8,50	7,90	7,00
Fc1c2[nH]ccc2cc(c1)N3CCNCC3Cc4ccccc4	7,80	8,20	7,80
C(C1CNCCN1c2cnc3[nH]ccc3c2)c4ccccc4	7,20	6,60	6,30
C(C1CNCCN1c2ccc3[nH]ncc3c2)c4ccccc4	8,90	7,60	7,80
C1CN(C(CN1)c2ccccc2)c3ccc4[nH]ncc4c3	7,70	6,70	6,70
C(Cc1ccccc1)C2CNCCN2c3ccc4[nH]ncc4c3	8,40	7,10	7,50
CCCC1CNCCN1c2ccc3[nH]ncc3c2	7,80	7,40	7,20
CCC[C@H]1CNCCN1c2ccc3[nH]ncc3c2	8,20	6,80	6,10
CCC[C@@H]1CNCCN1c2ccc3[nH]ncc3c2	7,50	7,50	7,60
CCCCC1CNCCN1c2ccc3[nH]ncc3c2	8,40	7,10	7,50

CC(C)CC1CNCCN1c2ccc3[nH]ncc3c2	8,30	7,70	7,80
C(C1CCCCC1)C2CNCCN2c3ccc4[nH]ncc4c3	7,40	5,80	6,50
COCC1CNCCN1c2ccc3[nH]ncc3c2	6,80	6,50	6,30
CCOCC1CNCCN1c2ccc3[nH]ncc3c2	7,80	6,50	6,50
COCCC1CNCCN1c2ccc3[nH]ncc3c2	8,70	6,60	6,80
C(C1CCOCC1)C2CNCCN2c3ccc4[nH]ncc4c3	8,40	6,50	7,40
C(C1CCOCC1)[C@H]2CNCCN2c3ccc4[nH]ncc4c3	9,00	6,50	6,00
C(C1CCOCC1)[C@@H]2CNCCN2c3ccc4[nH]ncc4c3	7,10	7,00	7,30
Fc1c2[nH]ccc2cc(c1)C(=O)C3(Cc4ccccc4)CCNC3	8,20	8,00	7,90
CCCC1(CCNC1)C(=O)c2cc(F)c3[nH]ccc3c2	8,50	8,20	7,80
CC(C)CC1(CCNC1)C(=O)c2cc(F)c3[nH]ccc3c2	9,00	8,60	8,10
CCOCC1(CCNC1)C(=O)c2cc(F)c3[nH]ccc3c2	8,50	7,70	7,10
CCCCC1(CCNC1)C(=O)c2cc(F)c3[nH]ccc3c2	8,90	8,60	8,20
CC(C)CCC1(CCNC1)C(=O)c2cc(F)c3[nH]ccc3c2	9,10	9,10	8,80
CC(C)(C)CCC1(CCNC1)C(=O)c2cc(F)c3[nH]ccc3c2	9,10	8,90	9,10
O=C(c1ccc2[nH]ccc2c1)C3(Cc4ccccc4)CCNC3	8,20	7,70	8,10
Cc1ccc(cc1Cl)C(=O)C2(CCC(C)(C)C)CCNC2	9,00	8,40	7,90
CC(C)(C)CCC1(CCNC1)C(=O)c2ccc(Cl)c(Cl)c2	8,60	8,20	8,20
CC(C)(C)CCC1(CCNC1)C(=O)c2cc(F)c(Cl)c(Cl)c2	8,50	8,00	8,00
CC(C)(C)CCC1(CCNC1)C(=O)c2cc(Cl)c(Cl)s2	8,30	8,10	8,70
CC(C)(C)CCC1(CCNC1)C(=O)c2ccc(Cl)c(Cl)n2	7,90	8,00	8,00
CC(C)(C)CCC1(CCNC1)C(=O)c2ccc(N)c(Cl)c2	9,00	8,30	8,50
CC(C)(C)CCC1(CCNC1)C(=O)c2cnc(N)c(Cl)c2	7,50	7,30	7,20
CC(C)(C)CCC1(CCNC1)C(=O)c2ccc3[nH]ncc3c2	9,20	8,20	8,10
CC(C)(C)CCC1(CCNC1)C(=O)c2cc3c(s2)cccc3	9,00	8,60	8,90
CC(C)(C)CCC1(CCNC1)C(=O)c2cc3c([nH]2)cccc3	8,20	8,60	8,60
CC(C)(C)CCC1(CCNC1)C(=O)c2ccc3ccccc3n2	8,70	8,50	8,40
Cc1cccc2nc(ccc12)C(=O)C3(CCC(C)(C)C)CCNC3	9,70	9,20	9,00
O=C(c1ccc2[nH]ccc2c1)C3(Cc4ccccc4)CCNCC3	8,10	7,00	8,00
O=C(c1ccc2[nH]ccc2c1)C3(Cc4ccccc4)CNC3	6,10	6,80	6,10
O=C(c1ccc2[nH]ccc2c1)C3(Cc4ccccc4)CCCNC3	7,90	7,60	7,20
O=C(c1ccc2[nH]ccc2c1)C3(Cc4ccccc4)CCCCNC3	7,10	6,60	6,60
CCCCC1(CCCN1)C(=O)c2ccc3[nH]ccc3c2	8,70	8,50	8,60
CCCCC1(CCCN1)C(=O)c2cc(F)c3[nH]ccc3c2	8,20	8,70	8,40
CCCCC1(CCCN1)C(=O)c2ccc(Cl)c(Cl)c2	8,60	7,70	7,80
CCCCC1(CCCN1)C(=O)c2ccc(N)c(Cl)c2	8,70	8,10	8,20
CC(C)(C)CCCC1(CCCN1)C(=O)c2cc3c([nH]nc3)cc2	8,30	8,10	8,50

Gustavo H. M Sousa

04/11/2020

Data e assinatura do aluno(a)

03/11/2020

Data e assinatura do orientador(a)