

# Desenvolvimento de um Modelo Preditivo para a Atividade Inibitória de Tiazolopiridinil-ureias sobre a Subunidade B da DNA Girase de Micobactérias por Meio de Descritores Moleculares

## *Development of a Predictive Model for Inhibitory Activity of Thiazolopyridinyl-ureas on Mycobacterial DNA Gyrase Subunit B using Molecular Descriptors*

Nathalia de Azevedo,<sup>a</sup> Felipe L. C. Silva,<sup>b</sup> Carlos M. R. Sant'Anna<sup>c,\*</sup>

<sup>a</sup> Universidade Federal Rural do Rio de Janeiro, Instituto de Ciências Exatas, Programa de Pós-Graduação em Modelagem Matemática e Computacional, Rodovia BR 465, km07, s/n, Zona Rural, CEP 23890-000, Seropédica-RJ, Brasil.

<sup>b</sup> Universidade Federal Rural do Rio de Janeiro, Instituto de Ciências Exatas, Departamento de Matemática, Rodovia BR 465, km07, s/n, Zona Rural, CEP 23890-000, Seropédica-RJ, Brasil.

<sup>c</sup> Universidade Federal Rural do Rio de Janeiro, Instituto de Química, Departamento de Química Fundamental, Rodovia BR 465, km07, s/n, Zona Rural, CEP 23890-000, Seropédica-RJ, Brasil.

\*E-mail: [santanna@ufrj.br](mailto:santanna@ufrj.br)

Recebido em: 14 de Outubro de 2021

Aceito em: 18 de Fevereiro de 2022

Publicado online: 7 de Abril de 2022

The increase in the incidence of resistant tuberculosis cases in recent years has stimulated the search for new drug candidates capable of acting on new molecular targets. *Mycobacterium tuberculosis*, which causes tuberculosis, has a single type II topoisomerase, DNA gyrase, responsible for maintaining the DNA topology during replication, transcription and recombination. One of the DNA gyrase subunits, the B subunit (GyrB), performs the hydrolysis of ATP (ATPase activity). In this work, correlation models were developed between activity data of known mycobacteria GyrB inhibitors with molecular descriptors obtained through molecular modeling calculations. To achieve this goal, molecular docking, semi-empirical and multiple linear regression methods were combined to build antimycobacterial activity prediction models (pIC50) for series of thiazolopyridinyl-ureas with inhibitory action on GyrB. The internal validation of the models was performed using the Leave-One-Out (LOO) cross-validation method. The results of the cross-validation were expressed by the correlation coefficient of the cross-validation ( $Q^2$ ) and the standard deviation of the cross-validation (SPRESS). The model's prediction statistics are expressed by the multiple correlation coefficient  $R^2_{ext}$  and the root mean square of the prediction error (RMSEP). Models with good statistical quality were obtained for a series of thiazolopyridinyl-ureas.

**Keywords:** *Mycobacterium tuberculosis*; molecular docking; molecular descriptors; linear regression

## 1. Introdução

O bacilo *Mycobacterium tuberculosis* é o principal agente causador da tuberculose (TB). Cerca de 10 milhões de pessoas desenvolveram essa doença e aproximadamente 1,4 milhão de pessoas vieram a óbito por causa da TB no ano de 2019 no mundo.<sup>1</sup> Quando disponível e usado corretamente, o tratamento com quimioterápicos é capaz de curar a quase totalidade dos casos de TB. No entanto, cepas de *M. tuberculosis* que apresentam resistência a um ou mais dos fármacos quimioterápicos de primeira escolha têm surgido em razão de mutações em diferentes genes bacterianos e da seleção dos indivíduos mais aptos dentre os submetidos à ação desses fármacos, podendo causar a chamada tuberculose multirresistente (TB-MDR, do inglês *MultiDrug-Resistant*).<sup>2</sup> Segundo a OMS, cerca de 465 mil pessoas foram diagnosticadas com TB-MDR no ano de 2019 e, destas, mais de 60% não conseguiram obter acesso ao tratamento.<sup>1</sup>

Essa situação traz risco quanto ao controle da TB no mundo e tem feito com que se busquem candidatos a fármacos capazes de atuar em novos alvos terapêuticos. A DNA girase é um importante alvo de ação de alguns fármacos usados no tratamento da TB, como as fluoroquinolonas, mas já há registro de cepas de *M. tuberculosis* resistentes a esta classe terapêutica. No entanto, as fluoroquinolonas atuam na subunidade A da DNA girase (GyrA), envolvida na clivagem do DNA, e a DNA girase contém outra subunidade, que realiza a hidrólise do ATP (GyrB). Em razão disso, novos inibidores da DNA girase, que apresentam um mecanismo de inibição alternativo na subunidade GyrB têm sido propostos.<sup>3</sup>

A modelagem molecular dispõe de diversos métodos que podem ser utilizados como ferramentas para auxiliar no planejamento racional de novos compostos bioativos.<sup>4</sup> Esses métodos podem ser usados, por exemplo, para a geração de descritores moleculares, números gerados por diferentes procedimentos que descrevem a estrutura molecular diretamente ou suas propriedades.<sup>5,6</sup> Atualmente, há programas que permitem o cálculo de uma extensa variedade de descritores, que incluem os físico-químicos, os topológicos, os termodinâmicos, os eletrônicos, entre outros.<sup>7</sup>

Descritores moleculares, juntamente com dados da atividade biológica de uma série de compostos, podem ser usados para se construir um modelo de correlação quantitativa entre a estrutura e a atividade biológica (QSAR, do inglês *Quantitative Structure Activity Relationship*).<sup>8</sup> São aplicados métodos matemáticos e estatísticos de análise de dados para se desenvolver modelos que estabelecem uma relação entre a atividade biológica e as estruturas químicas de uma série de compostos preferencialmente semelhantes, ou seja, compostos que se diferenciam pela presença de um ou mais grupos substituintes em posições estabelecidas na estrutura química comum à série.<sup>9</sup> Essa análise deve ser capaz de explicar o fenômeno observado e prever projeções a partir de um modelo que compara a medida quantitativa da atividade biológica Y (variável dependente) com m variáveis independentes X (descritores).<sup>10</sup> Esses descritores podem incluir propriedades físico-químicas da molécula como um todo, como sua massa molecular, refratividade molar,  $\log P$ , propriedades termodinâmicas e eletrônicas, como as energias dos orbitais de fronteira, ou de determinados fragmentos moleculares, como as constantes que descrevem os efeitos eletrônico (constantes de Hammett,  $s_m$  e  $s_p$ ) e estérico (fatores estéricos de Taft,  $E_s$ ) de substituintes; descritores que dependem de representações bidimensionais das moléculas, como os descritores topológicos; e descritores que descrevem propriedades tridimensionais e que dependem da conformação molecular, como a área de superfície acessível ao solvente e os descritores de campos moleculares, usados em métodos de QSAR-3D, como CoMFA e CoMSIA, entre outros.<sup>7</sup>

Um importante problema em QSAR-3D é a definição da conformação bioativa dos ligantes. Uma solução interessante é o uso de um método integrado, que inclui informações não somente do ligante, mas também do complexo receptor-ligante, método chamado rQSAR (QSAR baseado no receptor).<sup>10</sup> O rQSAR pode ser implementado usando-se,

por exemplo, docagem molecular (também conhecida como ancoramento, atracamento, ancoragem, acoplamento ou ainda pelo termo em inglês, *molecular docking*) como um meio para determinar as orientações e conformações mais prováveis dos ligantes localizados em alvos de estrutura conhecida para, em seguida, extrair-se descritores dos ligantes nessas “poses”, que podem ser aplicados no desenvolvimento de modelos QSAR em relação aos dados de bioatividade experimental.

Neste trabalho foram empregados métodos de modelagem molecular e de regressão linear múltipla para se construir modelos de predição de atividade inibitória de uma série de tiazolopiridinil-ureias com ação inibitória ( $pIC_{50}$ ) determinada sobre a subunidade GyrB da DNA girase.<sup>11</sup> Para se obter os modelos foram combinados descritores calculados tanto para os ligantes quanto para os complexos formados com a enzima.

## 2. Métodos

O esquema geral para construção de um modelo de QSAR é apresentado na Figura 1.<sup>12</sup>

O esquema geral descreve diversas etapas que vão desde a definição da variável y até finalmente se chegar a um modelo aceitável. Neste trabalho, a variável y são os valores de  $pIC_{50}$ , resultados experimentais da inibição da subunidade GyrB da DNA girase de *M. smegmatis*,<sup>11</sup> que apresenta elevada identidade com a enzima correspondente de *M. tuberculosis*, superior a 82%.

As outras etapas são descritas nos tópicos 2.1 e 2.2, a seguir. Deve-se destacar que obter um modelo de QSAR não é simplesmente obter uma equação matemática que correlacione os dados de forma estatisticamente aceitável; para ser um modelo de QSAR é preciso que as variáveis presentes na equação sejam consistentes em relação a um

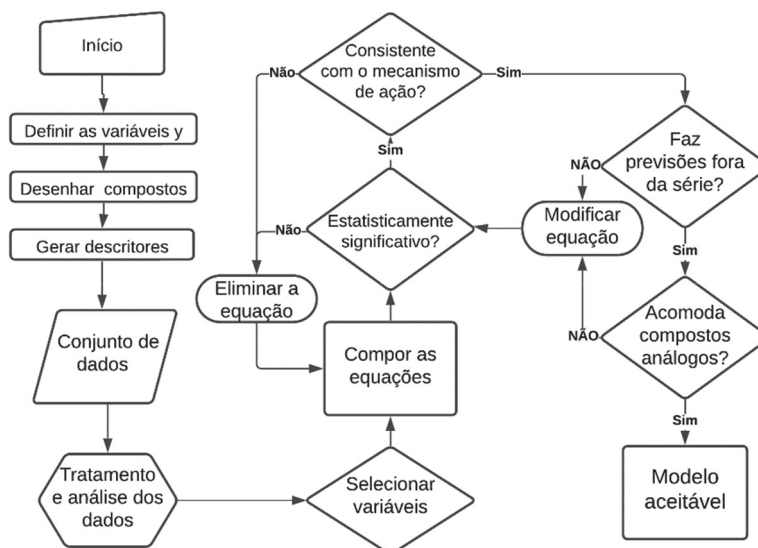


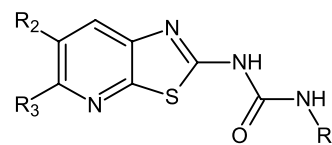
Figura 1. Esquema com as etapas para a construção de um Modelo de QSAR. Fonte: Adaptado de Gaudio e Zadonade, 2001<sup>12</sup>

proposto mecanismo de ação. Além disso, esse modelo deverá ser capaz de fazer previsões fora da série de compostos testada.<sup>12</sup>

Todas as análises estatísticas, gráficas e o desenvolvimento das estimativas de modelo desse trabalho foram realizadas no programa R, disponível em edições comerciais e de código aberto para diferentes plataformas (Windows, Mac e Linux).<sup>13</sup>

## 2.1. Estrutura molecular e otimização da geometria

A estrutura geral dos compostos usados neste trabalho é apresentada na Figura 2,<sup>11</sup> sendo os respectivos substituintes descritos na Tabela 1. Os compostos foram construídos através do programa Spartan'14 (Wavefunction, Inc.).<sup>14</sup> Este programa otimiza a geometria molecular da estrutura através da realização de cálculos quânticos e de mecânica molecular. As estruturas tiveram suas energias minimizadas usando o campo de força MMFF (*Merck Molecular Force Field*) e os mínimos obtidos foram refinados através de nova minimização de energia com o método quântico semi-empírico PM6,<sup>15</sup> sendo gravados em formato Sybyl (extensão mol2).



**Figura 2.** Estrutura geral das tiazolopiridinil ureias usadas neste estudo (vide Tabela 1 para a relação dos substituintes R<sub>1</sub>, R<sub>2</sub> e R<sub>3</sub>)

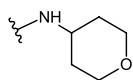
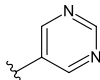
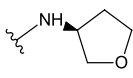
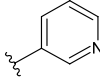
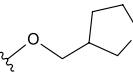
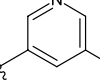
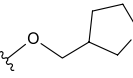
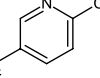
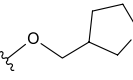
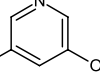
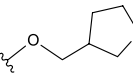
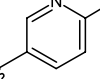
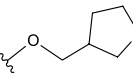
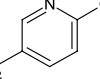
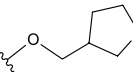
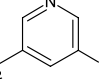
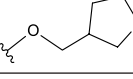
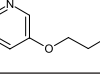
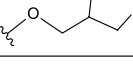
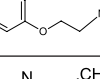
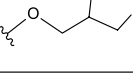
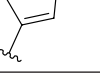
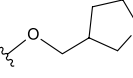
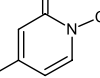
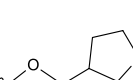
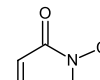

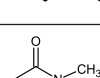
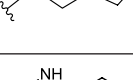
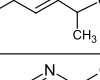
As atividades biológicas (variável dependente) estavam disponíveis como valores de IC<sub>50</sub> determinados sobre a subunidade GyrB da DNA girase de *M. smegmatis*.<sup>11</sup> Por se tratar de um dado experimental, para linearizar os dados foi aplicada a conversão para o negativo do logaritmo em base 10, pIC<sub>50</sub> = -log IC<sub>50</sub>, aos valores do IC<sub>50</sub> (Tabela 1).<sup>15</sup>

Foram avaliadas como variáveis independentes X (descritores moleculares) tanto descritores das moléculas isoladas calculados no programa Spartan quanto dos complexos correspondentes obtidos por docagem molecular, realizada no programa GOLD 5.7.3 (CCDC), que usa um algoritmo genético na busca de soluções.<sup>17</sup> Foi selecionada para a docagem a estrutura da GyrB de *M. smegmatis* depositada no *Protein Data Bank* (PDB) com código 4B6C, contendo o ligante co-cristalizado 6-(3,4-dimetilfenil)-3-[[4-[3-(4-metil-1-piperazinil)propoxi]fenil]amino]pirazina-2-

**Tabela 1.** Substituintes (vide Figura 2), variáveis dependentes e independentes selecionadas para os compostos usados neste estudo

Comp.	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	pIC <sub>50</sub>	LUMO <sup>a</sup>	S.vdw.ext <sup>b</sup>	S.hb.ext <sup>c</sup>
1	Etil	Metoxil		7,66	-1,5	51,91	51,39
2	Etil			8,43	-1,18	63,46	61,92
3	Alil	H		7,06	-1,36	52,92	51,54
4	Alil			7,34	-1,37	59,53	54,73
5	Etil			7,60	-1,34	57,38	55,71
6	Etil			7,92	-1,29	57,18	57,20
7	Etil			8,05	-1,31	57,07	54,74
8	Etil			7,89	-1,36	56,23	55,83
9	Etil			7,41	-1,43	54,91	54,55
10	Etil			7,77	-1,54	56,44	55,10

**Tabela 1.** Substituintes (vide Figura 2), variáveis dependentes e independentes selecionadas para os compostos usados neste estudo (cont.)

11	Etil			7,52	-1,01	58,95	57,94
12	Etil			7,52	-1,07	55,11	56,34
13	Etil			8,15	-1,49	57,3	56,18
14	Etil			8,70	-1,25	60,04	59,37
15	Etil			8,77	-1,38	58,73	57,98
16	Etil			9,05	-1,44	56,87	56,52
17	Etil			9,40	-1,36	62,18	61,88
18	Etil			9,22	-1,33	59,45	59,69
19	Etil			8,40	-1,29	64,76	66,73
20	Etil			9,00	-1,32	67,1	52,79
21	Etil			8,74	-1,09	56,79	57,26
22	Etil			9,85	-1,32	59,95	59,77
23	Etil			9,85	-1,33	62,12	62,46
24	Etil			9,28	-1,31	65,71	63,56
25	Etil			8,30	-0,94	64,66	65,63

<sup>a</sup> Energia do LUMO, método PM6, em eV. <sup>b</sup> Termo de van der Waals, associado às interações que ocorrem entre os átomos dos fragmentos apolares pertencentes ao ligante e à proteína, função GoldScore do programa de docagem molecular GOLD. <sup>c</sup> Termo de ligações de hidrogênio entre o ligante e o alvo, função GoldScore do programa de docagem molecular GOLD.

carboxamida, com resolução de 2,20 Å.<sup>18</sup> O raio da cavidade de interação foi de 15 Å, tendo como centro o Cg do resíduo Asp79. Para se selecionar a melhor função de pontuação para a docagem nesta estrutura, foi feito um processo de redocagem, no qual se avaliou, através dos valores de RMSD, qual das funções de pontuação disponíveis no programa GOLD (ASP, ChemPLP, ChemScore e GoldScore) seria a mais precisa na previsão do modo de interação do ligante co-cristalizado.

Ao todo, cerca de 30 descritores foram avaliados como variáveis independentes, incluindo descritores das moléculas isoladas (energia do HOMO, energia do LUMO, energia de solvatação) e descritores dos complexos (valores das pontuações e de todos os termos que compõe as pontuações das melhores poses, obtidos com as quatro funções de pontuação do programa GOLD); por exemplo, para a função GoldScore, os termos incluídos além da pontuação foram a energia da ligação de hidrogênio do complexo proteína-ligante (S.hb.ext), energia de van der Waals do complexo proteína-ligante (S.vdw.ext), energia de van der Waals interna do ligante (S.vdw.int), energia de tensão torcional do ligante (S.tor.int).

Antes de aplicar estratégias para selecionar quais variáveis deveriam ser incorporadas ao modelo, foi necessário determinar a matriz de covariância para avaliar a correlação entre as variáveis, para se evitar o uso de variáveis independentes correlacionadas. Para se definir quais descritores poderiam ser incorporados ao modelo, de modo a conseguir parâmetros de ajuste adequado sem haver o sobreajuste do modelo, foi usado o método de regressão para os melhores subconjuntos.<sup>19</sup> Esse método realiza combinações entre os possíveis subconjuntos das variáveis e cada possibilidade foi avaliada pelos parâmetros de ajuste  $R^2$  e  $S^2$ .<sup>19</sup>

## 2.2. Elaboração do modelo

A construção de um modelo de QSAR requer métodos de calibração multivariada, na qual o objetivo é estimar os valores apropriados dos coeficientes de regressão e prever uma estimativa realista dos erros de previsão.<sup>8</sup> As equações de regressão linear múltipla (RLM) são bastante usadas; entretanto, há algumas limitações, não sendo recomendadas em casos de variáveis altamente colineares e em casos que o número de objetos é menor que o número de variáveis regressoras.<sup>20</sup>

Verificou-se que o conjunto de dados desse trabalho se adequa ao método da RLM. Todos os modelos para os inibidores da GyrB de *M. smegmatis* passaram por testes para analisar se os resíduos assumem os pressupostos: seguem uma distribuição normal;<sup>21</sup> são independentes;<sup>22</sup> e têm variância  $\sigma^2$  constante (homocedasticidade).<sup>23</sup>

A validação interna de todos os modelos de QSAR foi realizada utilizando os compostos do conjunto treinamento, através do método de validação cruzada LOO (do inglês *Leave-One-Out*).<sup>24</sup> Esse método treina o modelo preditivo

com (n-1) observações de uma amostra de tamanho n. O procedimento foi repetido várias vezes até que todos os compostos do conjunto tivessem sido retirados e preditos ao menos uma vez.<sup>24</sup> Os resultados da validação cruzada foram expressos pelo coeficiente de correlação da validação cruzada ( $Q^2$ ) e pelo desvio-padrão da validação cruzada ( $S_{PRESS}$ ).<sup>12</sup>

A validação externa, dos modelos foi desenvolvida utilizando o conjunto teste contendo 30% dos compostos que não foram utilizados na geração dos modelos. As estatísticas de previsão do modelo foram expressas pelo coeficiente de determinação múltipla  $R_{ext}^2$ , também conhecido como coeficiente de correlação múltipla ou variância explicada,<sup>25</sup> e a raiz quadrada média do erro de previsão (RMSEP).<sup>25</sup> A avaliação estatística do modelo final considerou a relação em que o coeficiente de correlação  $R^2$ , calculado para o conjunto de treinamento, seja maior que a correlação da validação cruzada  $Q^2$ . Nos modelos de QSAR, o valor mínimo aceito para  $Q^2$  é 0,5, enquanto  $R^2 > 0,6$ . Caso a diferença entre  $R^2$  e  $Q^2$  exceda 0,2-0,3 isso é um forte indício de que o modelo teve um sobreajuste.<sup>25</sup>

A normalidade dos resíduos foi verificada pelo teste Shapiro-Wilk, cujo P-valor deve ser superior à 0,05 para indicar que possuem distribuição normal.<sup>21</sup> Diante do pressuposto que os resíduos seguem uma distribuição normal, foi realizado o teste de homoscedasticidade de Brush-Pagan, na qual, o P-valor também deve ser superior à 0,05 para indicar a homoscedasticidade.<sup>23</sup> O teste de Durbin-Watson foi utilizado para avaliar a independência dos resíduos e, assim, verificou-se que não são autocorrelacionados; os valores desse teste devem estar entre 1,5 e 2,5.<sup>22</sup>

## 3. Resultados e Discussão

Os resultados do processo de redocagem foram avaliados em função do valor de RMSD (Tabela 2). A função de pontuação, que gerou o melhor RMSD no sítio ativo da enzima (PDB: 4BC6), foi a GoldScore. No entanto, as outras funções apresentaram valores dentro do limite aceitável,  $RMSD < 2 \text{ \AA}$ .

**Tabela 2.** Valores de RMSD do processo de redocagem do ligante co-cristalizado na estrutura 4BC6, programa GOLD 5.7.3

Função	Melhor RMSD (Å)
ASP	1,299
ChemPLP	1,096
ChemScore	1,270
GoldScore	0,528

Sendo assim, os estudos posteriores de docagem molecular foram feitos com todas as 4 funções. Ao finalizar a docagem molecular, o programa GOLD 5.7.3 disponibiliza uma lista com a melhor pontuação de cada



estrutura analisada. Nessa lista, além da pontuação, estão discriminados todos os termos que compõem a pontuação de cada função específica. Como exemplo, para a função GoldScore existem dados referentes ao termo das ligações de hidrogênio entre a proteína e o ligante, definido pela sigla S(h bond). Nesse trabalho, os dados de todos os termos são parte dos descritores moleculares utilizados como  $m$  variáveis descritivas X do modelo.<sup>20,26</sup> Os cálculos foram feitos em triplicata; dessa forma, foi feita a média aritmética das três listas antes de se começar as análises.

No programa GOLD 5.7.3 é possível aplicar a técnica de rescore (repontuação), no processo de docagem. Essa técnica recalcula a pontuação de uma pose gerada por uma função com uma outra função. Diferentes funções de pontuação podem ter um desempenho melhor em casos específicos. Por exemplo, uma função pode superar a outra na classificação de compostos com um tipo de atividade, enquanto o inverso pode ocorrer com outras classes.<sup>17</sup> Nesse trabalho utilizou-se essa estratégia na tentativa de se obter diferentes pontuações e os parâmetros associados são parte das variáveis independentes. A docagem foi realizada com a função GoldScore e a repontuação com a função ChemPLP.

Todos os descritores resultantes da docagem molecular somados aos descritores dos ligantes isolados calculados com o programa Spartan foram explorados na busca de correlações com as atividades. Como exemplo do processo usado na obtenção dos modelos, das 25 observações da série das tiazolopiridinil ureias foi construído inicialmente o modelo 1, Equação 1, que não conduziu a um bom resultado de correlação,  $R^2 = 0,44$ . Como a técnica de RLM é bastante sensível a pontos de alavancagem e *outliers*,<sup>25</sup> foi realizada a análise gráfica (Figura 3) da Equação 1, para identificar as observações que estariam influenciando desfavoravelmente o modelo.

$$pIC50 = -1,369 - 0,945LUMO + 0,0423S.vdw_{ext} - 0,0490S.hb_{ext} \quad (1)$$

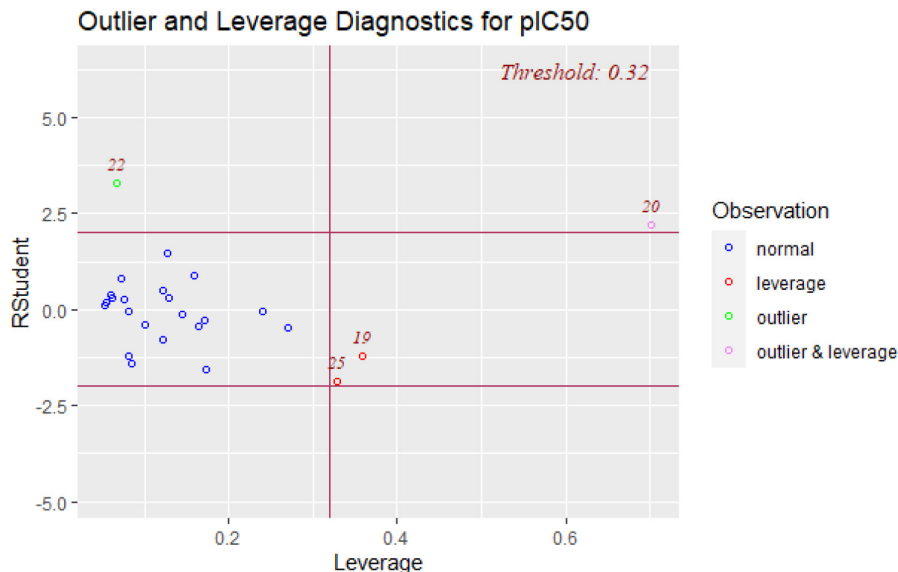
onde LUMO é o descritor da energia do LUMO, obtido com o método PM6 do programa Spartan'14; a função GoldScore é uma função de campo de força e  $S.vdw_{ext}$  é um descritor que representa o termo de van der Waals do complexo entre o ligante e a proteína, e  $S.hb_{ext}$  é o descritor do termo das ligações de hidrogênio entre o ligante e a proteína, obtido nos cálculos de docagem, utilizando a técnica de repontuação, ou seja, para a mesma geometria de interação obtida com a função GoldScore foi feita a repontuação com a função ChemPLP.

As observações indicadas no gráfico foram retiradas uma de cada vez, e em seguida elaborou-se os possíveis modelos. Após investigações, retirou-se os compostos **19**, **20**, **22** e **25**. Com esse novo conjunto de dados contendo 21 observações, chegou-se ao modelo 2, Equação 2.

$$pIC50 = -1,058 - 1,551LUMO + 0,070S.vdw_{ext} - 0,0911S.hb_{ext} \quad (2)$$

onde os descritores são os mesmos descritos no modelo 1.

O modelo 2 difere do modelo 1 no tamanho (n) da amostra, pois foram retirados os pontos influentes do conjunto de dados. A retirada desses pontos influentes do conjunto de dados foi capaz de melhorar significativamente os coeficientes de ajustes. O modelo 2 gerou a correlação com  $R^2 = 0,82$  para o conjunto de treinamento. A estatística de avaliação interna desse modelo foi verificada pelo  $Q^2 = 0,78$ ; e o desvio-padrão da validação cruzada,  $S_{PRESS}$ , foi 0,063. Na validação externa, obteve-se  $R^2 = 0,73$ ; e RMSEP = 0,042. O modelo foi aprovado pelos testes de análise de resíduos. O teste de Shapiro-Wilk teve valor de 0,13. O teste de Durbin-Watson teve valor de 2,49; e o teste de Brush-Pagan teve P-valor de 0,69.



**Figura 3.** Gráfico para detectar observações influentes do modelo 1. Resíduos estudentizados serão mais eficazes para detectar observações Y periféricas do que resíduos padronizados

### 3.1. Análises dos termos envolvidos no modelo

#### 3.1.1. Energia dos orbitais LUMO

A utilização das energias dos orbitais de fronteira HOMO e LUMO como descritores moleculares em QSAR se baseia na teoria dos Orbitais de Fronteira, desenvolvida por Fukui e colaboradores na década de 1950. Fukui observou a reatividade de compostos aromáticos e percebeu que é possível prever o local preferencial de ataque de um reagente nucleofílico, eletrofílico ou radicalar de uma molécula através da análise dos orbitais HOMO e LUMO sobre cada átomo das moléculas envolvidas na reação química.<sup>27,28</sup>

O LUMO é um orbital vazio, e o fato da energia desse orbital surgir como um descritor no modelo 2 é um indicativo de que ao menos para parte dos compostos selecionados a inibição da enzima pode envolver algum caráter eletrofílico, possivelmente atuando como inibidores que se ligam covalentemente à enzima. Os compostos estudados neste trabalho possuem grupos ureia e, de fato, há registros na literatura de inibidores enzimáticos da classe das ureias que causam a carbamoilação de resíduos do sítio ativo de enzimas, inativando-as.<sup>29</sup>

Assim como no trabalho de Li e colaboradores,<sup>30</sup> no modelo 2, o sinal negativo no coeficiente do LUMO indica que a atividade inibitória pode aumentar (maior  $pIC_{50}$ ) à medida que a energia do orbital LUMO diminui. Isso está de acordo com a teoria dos orbitais de fronteira, porque quanto menor for a energia do LUMO do inibidor, mais fácil seria um ataque nucleofílico da enzima sobre esse inibidor, que levaria finalmente à carbamoilação e inativação da enzima.

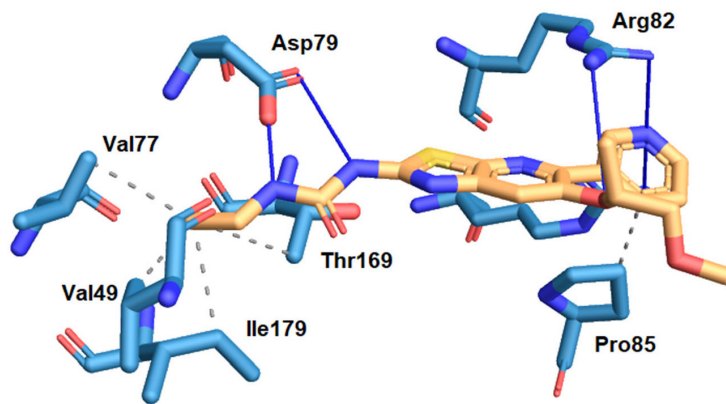
#### 3.1.2. Ligações de hidrogênio e interações de van der Waals entre ligante e proteína

A variável  $S_{Hbond}$  representa o termo da ligação de hidrogênio entre a proteína e o ligante da função GoldScore.<sup>31</sup> Ligações de hidrogênio são interações atrativas entre um átomo de hidrogênio de uma molécula ou fragmento molecular X-H, em que X é mais eletronegativo que H, e um átomo ou grupo de átomos na mesma

molécula ou em uma molécula diferente, sendo de grande relevância no planejamento de fármacos.<sup>32,33</sup> As ligações de hidrogênio ditas clássicas acontecem quando o átomo de hidrogênio está ligado a nitrogênio, flúor ou oxigênio e interage com um desses mesmos átomos na mesma ou em outra molécula.<sup>33</sup> Há de fato alguns grupos nas estruturas estudadas neste trabalho que participam diretamente de ligações de hidrogênio com resíduos do sítio da GyrB nos complexos obtidos na docagem, como o grupo ureia. Como exemplo, as principais interações observadas para o composto **6** são apresentadas na Figura 4, gerada com o servidor PLIP2021,<sup>34</sup> em que são destacadas as ligações de hidrogênio entre o grupo ureia e o resíduo Asp79 e entre o grupo 3-piridinil e o resíduo Arg82.

Em boa parte da literatura e no contexto do programa GOLD, as interações de van der Waals são descritas como interações que ocorrem entre os átomos de duas moléculas apolares.<sup>32</sup> A distância entre os núcleos dos átomos das espécies que interagem é que determina se essas forças são atrativas ou repulsivas e, por isso, é comum o uso de funções que somam um termo atrativo e um repulsivo para descrevê-las, como o potencial de Lennard-Jones. Quando a distância é muito curta, há uma sobreposição das nuvens eletrônicas que ocasiona no aumento de energia. Essa área estericamente repulsiva tende a diminuir a afinidade de ligação receptor-ligante, enquanto nas distâncias mais longas a tendência é que exista uma região de atração entre os átomos devido às forças de dispersão ocasionadas por interações favoráveis entre dipolos instantâneos e induzidos em dois átomos adjacentes. Essa força de interação é pequena, em média 0,2 kJ/mol, se for considerar um único contato átomo-átomo.<sup>32</sup>

No entanto, do ponto de vista da Química Medicinal, essas regiões de baixa ou nenhuma polaridade estão frequentemente associadas a grupos chamados hidrofóbicos, que ao interagir um com o outro deslocam moléculas de água de suas camadas de solvatação, resultando em um efeito entrópico favorável para interação.<sup>32</sup> Como exemplo, a Figura 4 apresenta as principais interações hidrofóbicas observadas para o composto **6**, em que são destacadas as interações entre o



**Figura 4.** Representação das principais interações entre o composto **6** e resíduos do sítio catalítico da GyrB da DNA girase. Ligações de hidrogênio são representadas em linhas contínuas azuis; interações hidrofóbicas como linhas pontilhadas cinza. Átomos de C do ligante são mostrados em laranja; átomos de C dos resíduos são mostrados em ciano; átomos de O, em vermelho; átomos de N, em azul; átomo de S, em amarelo.

grupo etil e os resíduos Val49, Val77, Thr169 e Ile171, e entre o grupo 3-piridinil e o resíduo Pro85.

#### 4. Conclusão

Neste trabalho foram combinados métodos de modelagem molecular e de regressão linear múltipla para se construir modelos de predição de atividade inibitória ( $pIC_{50}$ ) para uma série de 21 tiazolopiridinil ureias da literatura com ação inibitória sobre a subunidade B da DNA girase de *M. smegmatis*.

Diferente de estudos de QSAR tradicionais, em que se usam apenas descritores dos ligantes como variáveis independentes, aqui usamos o método de docagem molecular para se gerar também descritores relacionados à formação dos complexos enzima-inibidor. A docagem molecular foi usada para se prever o modo de interação mais provável entre cada inibidor e a enzima, e gerar uma série de descritores relacionados. Também foram usados cálculos quânticos semiempíricos para se gerar descritores dos ligantes, como a energia dos orbitais de fronteira. Combinando-se dados dos dois métodos, foi possível se obter correlações estatisticamente significativas, com coeficiente de ajuste  $R^2 = 0,82$ ; e o coeficiente de validação cruzada  $Q^2$  foi 0,78, entre os resultados de atividade inibitória da enzima e os descritores selecionados.

Esses modelos, calibrados por meio dos dados experimentais de atividade inibitória da enzima, podem servir como ferramentas para a predição de atividade de novas moléculas planejadas para agirem como inibidores da subunidade B da DNA girase, como candidatos a novos compostos com ação antimicobacteriana.

#### Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Os autores agradecem ao CNPq, INCT-INOVAR e FAPERJ pelo apoio financeiro concedido e bolsas concedidas.

#### Referências Bibliográficas

- World Health Organization. Global tuberculosis report 2020. World Health Organization: Geneva, 2020. Licence: CC BY-NC-SA 3.0 IGO. [\[Link\]](#)
- Rossetti, M. L. R.; Valim, A. R. D. M.; Silva, M. S. N.; Rodrigues, V. S.; Tuberculose resistente: revisão molecular. *Revista de Saúde Pública* **2002**, *36*, 525. [\[Crossref\]](#)
- Chaudhari, K.; Surana, S.; Jain, P.; Patel, H. M.; *Mycobacterium tuberculosis* (MTB) GyrB inhibitors: An attractive approach for developing novel drugs against TB, *European Journal of Medicinal Chemistry* **2016**, *124*, 160. [\[Crossref\]](#) [\[PubMed\]](#)
- Sant'Anna, C. M. R.; Métodos de Modelagem Molecular para Estudo e Planejamento de Compostos Bioativos: Uma Introdução. *Revista Virtual de Química* **2009**, *1*, 49. [\[Crossref\]](#)
- Consonni, V.; Todeschini, R.; Em Recent Advances in QSAR Studies. Challenges and Advances in Computational Chemistry and Physics; Puzyn, T.; Leszczynski, J.; Cronin, M., eds.; Springer: Dordrecht, 2010. [\[Crossref\]](#)
- Sant'Anna, C. M. R.; Em Química Medicinal: As Bases Moleculares da Ação dos Fármacos, 3a. ed.; Barreiro, E. J.; Fraga, C. A. M.; ArtMed: Porto Alegre, 2015. [\[Crossref\]](#)
- Xue, L.; Bajorath, J.; Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial Chemistry & High Throughput Screening* **2000**, *3*, 363. [\[Crossref\]](#) [\[PubMed\]](#)
- Tropsha, A.; Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics* **2010**, *29*, 476. [\[Crossref\]](#) [\[PubMed\]](#)
- Alves, V. M.; Braga, R. C.; Muratov, E. N.; Andrade, C. H.; Quimioinformática: uma introdução. *Química Nova* **2018**, *41*, 202. [\[Crossref\]](#)
- Lushington, G. H.; Guo, J. X.; Wang, J. L.; Whither combine? New opportunities for receptor-based QSAR. *Current Medicinal Chemistry* **2007**, *14*, 1863. [\[Crossref\]](#) [\[PubMed\]](#)
- Kamsri, P.; Punkvang, A.; Hannongbua, S.; Suttisintong, K.; Kittakoop, P.; Spencer, J.; Pungpo, P.; In silico study directed towards identification of the key structural features of GyrB inhibitors targeting MTB DNA gyrase: HQSAR, CoMSIA and molecular dynamics simulations. *SAR and QSAR in Environmental Research* **2019**, *30*, 775. [\[Crossref\]](#) [\[PubMed\]](#)
- Gaudio, A. C.; Zandonade, E. P.; Validação e Análise dos Modelos que Correlacionam Estrutura Química e Atividade Biológica. *Química Nova* **2001**, *24*, 658. [\[Crossref\]](#)
- R Core Team; R: A language and environment for statistical computing. R Foundation for Statistical Computing **2020**, Vienna, Austria. [\[Link\]](#)
- Hehre, W. J.; Deppmeier, B. J.; Klunzinger, P. E.; Wavefunction Inc. Irvine, California. 1999. [\[Link\]](#)
- Stewart, J. J.; Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling*, **2007**, *13*, 1173. [\[Crossref\]](#) [\[PubMed\]](#)
- Ferreira, R. S.; Glaucius, O.; Andricopulo, A. D.; Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala: oportunidades e desafios em P&D de fármacos. *Química Nova* **2011**, *34*, 1770. [\[Link\]](#)
- CCDC GOLD-Protein Ligand Docking Software. [\[Link\]](#)
- Shirude, P. S.; Madhavapeddi, P.; Tucker, J. A.; Murugan, K.; Patil, V.; Basavarajappa, H.; Raichurkar, A. V.; Humnabdar, V.; Hussein, S.; Sharma, S.; Ramya, V. K.; Narayan, C. B.; Balganes, T. S.; Sambandamurthy, V. K.; Aminopyrazinamides: Novel and specific GyrB inhibitors that kill replicating and nonreplicating *Mycobacterium tuberculosis*. *ACS Chemical Biology* **2013**, *8*, 519. [\[Crossref\]](#) [\[PubMed\]](#)
- Hofmann, M.; Gatu, C.; Kontoghiorghes, E. J; Efficient algorithms for computing the best subset regression models for



- large-scale problems. *Computational Statistics & Data Analysis* **2007**, *52*, 16. [[Crossref](#)]
20. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Wasserman, W.; Applied Linear Statistical Models, 5a. ed., Boston: McGraw-Hill, 2006. [[Link](#)]
21. Steinskog, D. J.; Tjøstheim, D. B.; Kvamstø, N. G.; A cautionary note on the use of the Kolmogorov–Smirnov test for normality. *Monthly Weather Review* **2007**, *135*, 1151. [[Crossref](#)] [[PubMed](#)]
22. Baum, C.; Wiggins, V.; BPAGAN: Stata module to perform Breusch-Pagan test for heteroskedasticity. **1999**. [[Link](#)]
23. TILLMAN, J. A.; The power of the Durbin-Watson test. *Econometrica* **1975**, *43*, 959. [[Crossref](#)]
24. Wong, T. T.; Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* **2015**, *48*, 2839. [[Crossref](#)]
25. Kiralj, R.; Ferreira, M.; Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *Journal of the Brazilian Chemical Society* **2009**, *20*, 770. [[Crossref](#)]
26. Varmuza, K.; Filzmoser, P.; Introduction to multivariate statistical analysis in chemometrics. CRC Press, 2016. [[Crossref](#)]
27. Fukui, K.; Yonezawa, T.; Shingu, H.; A molecular orbital theory of reactivity in aromatic hydrocarbons. *Journal of Chemical Physics* **1952**, *20*, 722. [[Crossref](#)]
28. Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H.; Molecular orbital theory of orientation in aromatic, heteroaromatic, and other conjugated molecules. *Journal of Chemical Physics* **1954**, *22*, 1433. [[Crossref](#)]
29. Johnson, D. S.; Stiff, C.; Lazerwith, S. E.; Kesten, S. R.; Fay, L. K.; Morris, M.; Ahn, K.; Discovery of PF-04457845: A Highly Potent, Orally Bioavailable, and Selective Urea FAAH Inhibitor. *ACS Medicinal Chemistry Letters* **2011**, *2*, 91. [[Crossref](#)] [[PubMed](#)]
30. Li, B. J.; Chiang, C. C.; Hsu, L. Y.; QSAR Studies of 3,3'-(Substituted-Benzylidene)-Bis-4-Hydroxycoumarin, Potential HIV-1 Integrase Inhibitor. *Journal of the Chinese Chemical Society* **2010**, *57*, 742. [[Crossref](#)]
31. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D.; Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52*, 609. [[Crossref](#)] [[PubMed](#)]
32. Fokoue, H. H.; Pinheiro, P. S. M.; Fraga, C. A. M.; Sant'Anna, C. M. R.; Há algo novo no reconhecimento molecular aplicado à química medicinal? *Química Nova* **2020**, *43*, 78. [[Crossref](#)]
33. Krogsgaard-Larsen, P.; Strømgaard, K.; Madsen, U.; Textbook of drug design and discovery, 4a. ed.; CRC Press: Boca Raton, 2002. [[Crossref](#)]
34. Adasme, M. F.; Linnemann, K. L.; Bolz, S. N.; Kaiser, F.; Salentin, S.; Haupt, V. J.; Schroeder, M.; PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Research* **2021**, *49*, W530. [[Crossref](#)] [[PubMed](#)]