Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Alexandre de Fátima Cobre

Desenvolvimento de modelos de machine learning baseados em QSAR-3D para predição de novos candidatos a fármacos inibidores da proteina CCR-5, para o tratamento de HIV/AIDS

Curitiba
2023

Alexandre de Fátima Cobre

Desenvolvimento de modelos de machine learning baseados em QSAR-3D para predição de novos candidatos a fármacos inibidores da proteina CCR-5, para o tratamento de HIV/AIDS

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Anderson Ara

Curitiba
2023

Alexandre de Fátima Cobre, Roberto Pontarolo, Anderson Ara*2, PhD., Corresponding author: Department of Statistics, Universidade Federal do Paraná, Av. Cel. Francisco H. dos Santos, 100 - Jardim das Américas, Curitiba city - PR, Zip code 81530-000, Brazil. E-mail: ara@ufpr.br. Telephone number: +55 41 88236116.

DSBD Especialização em Data Science & Big Data
dsbd.leg.ufpr.br

Universidade Federal do Paraná
Trabalho de Conclusão de Curso, nº 01, 01 (2023)

Desenvolvimento de modelos de machine learning baseados em QSAR-3D para predição de novos candidatos a fármacos inibidores da proteína CCR-5, para o tratamento de HIV/AIDS

Predição de novos candidatos a fármacos inibidores da proteína CCR-5, para o tratamento de HIV/AIDS

Alexandre de Fátima Cobre[1], Roberto Pontarolo[2], Anderson Ara[3]

[1]Programa de pós-graduação em Data Science Big Data, Departamento de Informática, Universidade Federal do Paraná, Curitiba, PR, Brasil*
[2]Departamento de Farmácia, Universidade Federal do Paraná, Curitiba, PR, Brasil†
[3]Programa de pós-graduação em Data Science Big Data, Departamento de Informática, Universidade Federal do Paraná, Curitiba, PR, Brasil‡

Introdução. C-C receptor quimiocina tipo 5 (CCR-5), é uma proteína encontrada na superfície das células de defesa (linfócitos e macrófagos). A CCR-5 é a estrutura à qual o vírus HIV (vírus da imunodeficiência humana) se liga para invadir a célula hospedeira causando o desenvolvimento da AIDS (síndrome da imunodeficiência adquirida). Neste estudo, foram desenvolvidos modelos de machine learning (ML) baseados em relação estrutura atividade quantitativa (QSAR) para predizer compostos com bioatividade inibitória contra a proteína CCR-5 para o tratamento de HIV. **Material e métodos**. Um conjunto de dados experimentais não reduntantes de 2929 compostos com valores de bioatividade inibitória (expressa em IC50) contra a proteína CCR-5 foram colectados na base de dados CHEMBL e empregados para desenvolver modelos de ML baseados em QSAR, visando predizer a sua bioatividade. Esses 2929 compostos foram descritos usando Pubchem fungreprint e 32 diferentes algorítmos de ML foram treinados e testados. A avaliação do desempenho dos modelos foi feita utilizando as métricas R2, MSE, RMSE, MAE e tempo de treinamento. Cada um dos cinco melhores modelos de ML foi aplicado o método SHAP value visando identificar as features (descritores) mais importantes na predição da bioatividade dos compostos contra HIV. **Resultados**. Os cinco melhores modelos de ML que tiveram melhor desempenho na predição da bioatividade inibitória contra a proteína CCR-5 para o tratamento de HIV foram: Random Forest (RF), Histogram Gradient Boosting (HGBM), LGBM, Bagging e KNN, cujos valores de capaciadde preditiva (R2) variaram entre 82-87%. **Conclusão**. Neste estudo, foram desenvolvido cinco modelos de ML (RF, HGBM, LGBM, Bagging e KNN) para predizer a bioatividade inibitória dos compostos contra a proteína CCR-5 para a descoberta de novos fármacos contra HIV. Esses modelos de ML podem ser usados como um filtro de seleção de novas moléculas, que podem ser testadas nos experimentos in vitro e in vivo que visam a descoberta de novos fármacos inibidores da proteína CCR-5 para o tratamento potencial de HIV. .

Palavras-chave: HIV, CCR-5, Drug discovery, QSAR, Machine learning, ensaio in silico

Introduction. C-C chemokine receptor type 5 (CCR-5) is a protein found on the surface of defense cells (lymphocytes and macrophages). CCR-5 is the structure to which the HIV virus (human immunodeficiency virus) binds to invade the host cell causing the development of AIDS (acquired immunodeficiency syndrome). In this study, machine learning (ML) models based on quantitative structure activity relationship (QSAR) were developed to predict compounds with inhibitory bioactivity against the CCR-5 protein for the treatment of HIV. **Material e métodos**. A non-redundant experimental dataset of 2929 compounds with inhibitory bioactivity values (expressed in IC50) against the CCR-5 protein were collected from the CHEMBL database and used to develop QSAR-based ML models to predict their bioactivity. These 2929 compounds were described using PubChem fingerprint and 32 different ML algorithms were trained and tested. The evaluation of the performance of the ML models was made using the metrics R2, MSE, RMSE, MAE and training time. Each of the five best ML models was applied the SHAP values method to identify the most important features (descriptors) in predicting the bioactivity of compounds against HIV. Results: The five best ML models that had the best performance in predicting the inhibitory bioactivity against the CCR-5 protein for the treatment of HIV were:

Random Forest (RF), Histogram based Gradient Boosting (HGBM), LGBM, Bagging and KNN, whose predictive capacity values (R2) ranged between 82-87**Results**. The five best ML models that had the best performance in predicting the inhibitory bioactivity against the CCR-5 protein for the treatment of HIV were: Random Forest (RF), Histogram based Gradient Boosting (HGBM), LGBM, Bagging and KNN, whose predictive capacity values (R2) ranged between 82-87%. **Conclusion**. In this study, five ML models (RF, HGBM, LGBM, Bagging and KNN) were developed to predict the inhibitory bioactivity of compounds against the CCR-5 protein for the discovery of new drugs against HIV. These ML models can be used as a selection filter for new molecules, which can be tested in in vitro and in vivo experiments aimed at discovering new CCR-5 protein inhibitor drugs for the potential treatment of HIV.

1. Introdução

A pandemia de HIV/AIDS (vírus da imunodeficiência humana/sídrome da imunideficiência adquirida) continua sendo um problema de saúde pública de alta letalidade, apesar dos grandes avanços conquistados no campo da terapêutica medicamentosa, diagnóstica e dos programas sociais. Os dados recentes da Organização Mundial da saude (OMS), mostram que em 2021 foram registrados, 38,8 milhões de pessoas que vivam com HIV, 1,5 milhões de novos casos e destes 650 mil foram fatais [1].

A proteína CCR-5 é um receptor de quimiocina encontrado nas células do sistema imunológico humano, incluindo os linfócitos TCD4+. O vírus da imunodeficiência humana (HIV) usa o receptor CCR-5 como um co-receptor para entrar nas células do hospedeiro [2].

Pessoas que são geneticamente deficientes na produção do receptor CCR-5 têm uma maior resistência à infecção pelo HIV [3]. Essa descoberta levou ao desenvolvimento de medicamentos antirretrovirais que visam bloquear o CCR-5 e impedir que o HIV entre nas células do hospedeiro [3, 4, 5].

Um exemplo de medicamento antirretroviral que atua como um antagonista do CCR-5 é o maraviroc, que foi aprovado para uso clínico em 2007. O maraviroc é geralmente utilizado em combinação com outros medicamentos antirretrovirais como parte do tratamento da infecção pelo HIV em pacientes adultos [4].

O uso de medicamentos que visam o CCR-5 é uma abordagem promissora para o tratamento e prevenção da infecção pelo HIV, mas a resistência ao medicamento pode ocorrer. Além disso, é importante ressaltar que esses medicamentos não são uma cura para a infecção pelo HIV e devem ser usados em combinação com outros medicamentos antirretrovirais [5].

Machine learning tem um papel fundamental na descoberta de novos fármacos, pois permite analisar grandes conjuntos de dados de maneira eficiente e identificar padrões que seriam impossíveis de se detectar manualmente. Além disso, o machine learning pode ajudar a prever a eficácia de novas moléculas candidatas, bem como sua segurança e toxicidade [6–11].

Existem várias maneiras pelas quais o machine learning pode ser aplicado na descoberta de fármacos. Uma delas é modelagem molecular por QSAR (Quantitative Structure-Activity Relationship), que é uma técnica usada na descoberta de fármacos que busca estabelecer uma relação matemática entre a estrutura química de uma molécula e sua atividade biológica. A QSAR utiliza uma série de descritores moleculares para quantificar a estrutura química da molécula e, em seguida, emprega técnicas estatísticas para modelar a relação entre esses descritores e a atividade biológica [12, 13].

O machine learning pode ser utilizado para criar modelos QSAR mais precisos e eficazes. Em particular, o uso de algoritmos de aprendizado de máquina, como redes neurais e árvores de decisão, tem sido eficaz na previsão da atividade biológica de novas moléculas candidatas. O uso de técnicas de machine learning em QSAR tem várias vantagens em relação aos métodos tradicionais. Em primeiro lugar, o machine learning permite a análise de um grande número de descritores moleculares, o que pode ajudar a identificar as características mais importantes da molécula que influenciam a atividade biológica. Em segundo lugar, o machine learning pode ajudar a identificar interações não lineares entre os descritores moleculares e a atividade biológica, o que pode ser difícil a detecção usando métodos estatísticos tradicionais. Além disso, o uso de técnicas de machine learning em QSAR permite a criação de modelos mais precisos e robustos, o que pode ajudar a acelerar o processo de descoberta de fármacos. Isso é particularmente importante no desenvolvimento de novos tratamentos para doenças raras e emergentes, onde os recursos e o tempo são limitados. Assim, o uso de machine learning em QSAR é uma abordagem promissora para a descoberta de fármacos, pois per-

*alexandrecobre@gmail.com
†pontarolo@ufpr.br
‡ara@ufpr.br

mite a análise eficiente de grandes conjuntos de dados e a identificação de padrões complexos que seriam difíceis de detectar usando métodos tradicionais [9, 14–18].

Existem vários fármacos que foram descobertos usando abordagens QSAR e machine learning, por exemplo, Raltegravir - um inibidor da integrase usado no tratamento de HIV, Cetuximab - um anticorpo monoclonal usado no tratamento de câncer de cólon e cabeça e pescoço, Valsartana - um inibidor da enzima conversora de angiotensina usado no tratamento de hipertensão e insuficiência cardíaca, Sitagliptina - um inibidor da dipeptidil peptidase-4 usado no tratamento de diabetes tipo 2 e Ibrutinibe - um inibidor da tirosina quinase usado no tratamento de câncer de leucemia linfocítica crônica [19]. Neste contexto, o objetivo deste estudo, foi desenvolver modelos de machine learning baseados em abordagem QSAR visando a predição de compostos com bioatividade inibitória da proteína CCR-5 para o tratamento de HIV.

2. Material e métodos

Na Figura 1 é mostrado o fluxo de execução deste estudo. Em sumo, foram desenvovidos modelos de machine learning baseados em abordagem de relação quantitativa estrutura atividade (QSAR), visando predizer e analisar compostos bioativos inibidores da proteína CCR-5 humano, para o tratamento potencial de HIV/AIDS. O estudo foi realizado utilizando o guia da Organization for Economic Cooperation and Development (OECD), que inclui os seguintes passos: (i) um conjunto de dados com um ponto final; (ii) uma análise exploratória desses dados; (iii) utilizando um conjunto de diferentes algoritmos de machine learning supervisionados ; (iv) utilização de métricas de avaliação do desempenho dos modelos de machine learning e (v) interpretação mecanística do modelos de machine learning (análise das features importantes).

2.1. Descrição do banco de dados utilizados

Chembl database

O conjunto de dados utilizados nesse estudo para o desenvolvimento de modelos QSAR machine leafning para predição de compostos com bioatividade contra a proteína CCR-5 humana para o tratamento de HIV foram coletados no banco de dados ChEMBL 32 (https://www.ebi.ac.uk/chembl/). ChEMBL database é um banco de dados Britânico de domínio público que contém dados de bioatividade de mais de 2,4 milhões

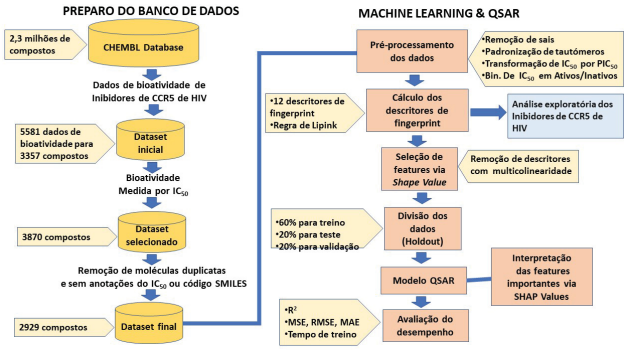

Figura 1: Fluxograma usado para coleta e limpeza de dados, e treinamento e validação dos modelos de machine learning para predição de bioatividade de compostos contra a proteína CCR-5 visando a descoberta de novos candidatos a fármacos tratamento de HIV.

de compostos com propriedades drug-like, distribuídos em 211 datasets. Chembl database, reune informações químicas dos compostos, dados de bioatividade, dados genômicos, com vista a auxiliar na tradução de informações genômicas em novos fármacos. O chembl database, também contém cerca de 86 mil publicações científicas, 1.5 milhões de ensaios, 15 mil alvos terapêuticos, 6.7 mil mecanismos de ação dos fármacos, 2 mil células, 43 mil indicações de medicamentos e 759 tecidos biológicos. Importante destacar que todas essas informações são selecionados manualmente e de forma curada por especialistas da área.

Dataset usado para o desenvolvimento do estudo

Na Chembl database, foram compilados compostos bioativos inibidores da proteína CCR-5 do HIV (Target ID CHEMBL274), que era composto po um total de 5581 dados de bioatividade correspondentes a 3357 compostos. Todas as notações smiles dos compostos foram curadas utilizando o ChemAxon's Standardizer, conforme descrito por Simeon (2016)[20] O dataset inicial continha diversas parametros de bioatividade tais como, EC50, MIC, porcentagem da atividade, Ki, porcentagem de inibição, IC50. O IC50 (medido na escala de nM) foi selecionado para posterior investição, pois é o parâmetro de bioatividade que estava disponível na maioria dos compostos (n = 2.929 compostos). Após, foi feita a remoção de compostos duplicados, compostos sem anotações do IC50 ou compostos que não apresentavam código SMILES. Nesta análise, nenhum composto foi removido, pois todos apresentavam dados completos. Assim o dataset final foi formado por 2.929 compostos bioativos, que foram estes utilzados para a próxima etápa doe estudo: o pré-processamento dos dados.

2.2. Pré-processamento dos dados

O pré-processamento dos dados de todos 2.929 compostos bioativos inibidores da proteína CCR-5 para o tratamento de HIV incluídos no estudo, foi realizado utilizando a biblioteca Padelpy da linguagem python (https://pypi.org/project/padelpy/). Este processo consistiu na remoção de sais, padronização de tautômeros, conversão do IC50 para pIC50 (a escala logaritmica). Foi também realizada a categorização dos valores de IC50 em três grupos de compostos: compostos ativos: IC50<100nM; compostos com atividade intermediária: IC50 entre 100-1000 nM; compostos inativos, IC50>1000nM[20]. Após essa etapa realizou-se o calculo dos descritores de fingerprint das moleculas. Eles consistem um conjunto de códigos binários que descrevem o espaço químico (2D, 3D, 4D, 5D ou até 6D) das moléculas. Existem 12 diferentes tipos de descritores de fingerprint (por exemplo, Pubchem, CDK, substructure, etc)[21]. Neste estudo, o fingerprint das moléculas foi calculado usando o descritor de Pubchem, que é uma representação binária de substruturas definidas por Pubchem. Para todo e qualquer molecula, existem 881 descritores (features) de Pubchem. Outros descritores adicionais foram tambem calculados con vista a definir a regra dos cinco de Lipinski (tambem conhecida como a regra da farmacêutica Pfizer), que comprendem: o peso molecular (MW), número de ligação aceptoras de hidrogênio (nHBAcc), número de ligações doadoras de hidrogênio (nHBDon) e logarítmo do coeficiente de partição octanol/água (LogP) [20]. A pfizer preconiza que para que para qualquer composto quimico tenha caracteristicas de um farmaco, é neccesario que os valores de MW500 g/mol; AlogP<5; nHBDon <10 e nHBAcc<5[22]. Esses descritores foram calculados utilziando a biblioteca RDKIT da linguagem python (https://www.rdkit.org/).

2.3. Feature selection

A multicolinearidade é um problema sério em análises de regerssão e é definida como sendo a intercorrelação entre pares de descritoers (features), que causa o aumento do viés e complexidade do modelo de machine learning desenvolvido. Visando a resolver este problema, foi aplicado um filtro de variância, onde os descritores (features) com variância <0,1 foram removidas do dataset, objetivando obter um subconjunto reduzido de descritores de Pubchem[23]. Todo esse processo foi programado em linguagem python (Fig. 1).

- **Data spliting**

O método Holdout foi empregado para a divisão dos dados para machine learning, com vista a minimizar o viés da seleção dos dados de treinamento, de teste e de validação (ver figura 1) [24]. Neste método, os dados foram divididos em três etapas: 60% para o treinamento do modelo, 20% previsão e para testagem do modelo [25]. Todo processo de divisão dos dados foi realizado utilizando um random state = 100.

- **Machine learning**

Nesta etapa do estudo, todas as análises foram realizadas em linguagem python. Modelos de regressão foram implementados para predizer a atividade biológica (expressa na forma de pIC50) a partir da estrutura química dos compostos bioativos (expressos na forma de descritores de Pubchem). Nesta análise a variável resposta foi o pIC50, ao passo que as variáveis preditoras foram os descritores de fingerprint de Pubchem. A biblioteca Scikit-Learn e LazyPredict foi utilizada para construir 32 diferentes algoritmos de machine learning (ML), visando selecionar os cinco modelos de ML com melhor desempenho preditivos. As seguintes métricas foram usadas para avaliação dos modelos de machine learning desenvolvidos: coeficiente de determinação (R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), mean absolute error (MAE) [26, 27]. O R2 é uma métrica que mede a capacidade preditiva do modelo de regressão. Os valores de R2 variam entre zero a um, e quanto mais próximo de um, mais preditivo é o modelo. Por outro lado, MAE, MSE e RMSE são métricas de erros. Um modelo de regressão bem ajustado, é aquele que busca minizar os valores de MAE, MSE e RMSE.

- **Interpretação mecanística das features importantes que contribuem na atividade biológica via SHAP values**

A interpretação das features importantes dos modelos de machine learning foi feita utilizando o método SHAP values. Para executar o SHAP values, inicialmente foi criado o objeto “explainer”, que ajuda a analisar um composto isolado ou um conjunto de compostos bioativos como um todo. Apóa, o próximo passo foi calcular os valores médios de cada feature usando a função softmax. O efeito global dos descritores de fingerprnt (features) foi analisada por meio do SHAP plot, beeswarm plot, summary plot, violine plot. Por outro

lado, para investigar quais são as partes de uma molécula específica (features) são importantes (ou não) para atividade biologica (pIC50) foi construído o Bar plot e Waterfall plot[28].

3. Resultados e discussão

3.1. Análise do espaço químico dos descritores

O espaço químico dos descritores dos compostos bioativos inibidores da proteína CCR-5 para o tratamento de HIV foi feita visando obter insight da relação estrutura atividade quantitativa (QSAR) usando os quatro descritores (AlogP, MW, nHBDon and nHBAcc) da regra dos 5 de Lipinski [29]. O AlogP é um parâmetro que define a medida de absorção de fármacos pelas membranas celulares. Quanto menor o AlogP, melhor será a absorção do fármaco através da bicamada lipídica das membranas [30, 31]. O MW é um parâmetro que também está associado a absorção dos fármacos, e quanto menor é o peso molecular de um fármaco mais rápida será a sua absorção [32, 33]. nHBDon e nHBAcc são usados para expressar o número de grupos doadores e aceptores de ligações de hidrogênio. Em geral, quanto maior o número de ligações de hidrogênio entre um fármaco e o seu receptor, maior é a sua bioatividade [34, 35].

A Figura 2 mostra os resultados da análise exploratória dos dados usando os descritores de Lipinski. Podemos observar que, os valores de pIC50 dos compostos ativos e inativos foram estatisticamente diferentes, para os compostos ativos e inativos contra a proteína CCR-5 humana (p <0,001). O threshold dos valores de pIC50 entre os dois grupos de compostos foi de 4,5, onde compostos ativos apresentam pIC50>4,5 e compostos inativos, pIC50<4,5. Outras diferenças observadas entre os dois grupos de compostos (ativos e inativos), foi em relação ao peso molecular, AlogP e grupos doadores de ligações de hidrogênio (nHBDon). No entanto, não foi observado nenhuma diferença significativa entre os dois grupos de compostos em relação ao grupo aceptores de ligações de hidrogênios. Nós também observamos uma correlação positiva entre o peso molecular (MW) e AlogP. Esses resultados mostram que realmente o espaço químico entre os compostos com bioatividade e sem bioatividade contra a proteína CCR-5 humana para o tratamento de HIV, são realmente diferente. Esses resultados, são também condizentes com a literatura [36–38]. Esses insights serviram de justificativa para elaboração de modelos de machine learning baseados em QSAR para investigação de potenciais fármacos inibidores da proteína CCR-5 para o tratamento de HIV.

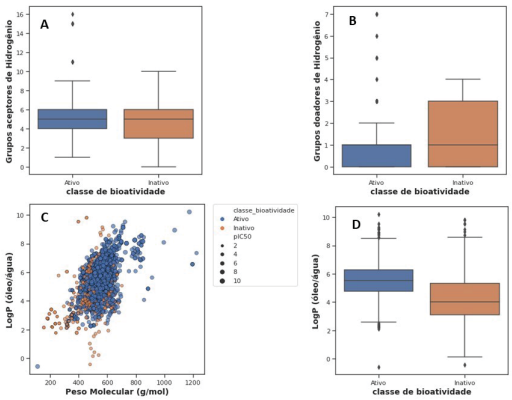

Figura 2: Resultados de análise exploratória dos compostos analisados usando os descritores de Lipinski.

- **Machine learning QSAR**

Um total de trinta e dois (32) diferentes algoritmos de machine learning foram treinados e validados (figuras S1-S2, em material suplementar), visando predizer a atividade biológica (pIC50) a partir dos descritores moleculares de Pubchem. Importante destacar que dos 881 descritores de Pubchem calculados[39], apenas 204 deles foram selecionados para o treinamento dos modelos de machine learning, o que permitiu minimizar os problemas de multicunolinearidades nos modelos de machine learning treinados [40, 41]. Dos 32 algoritmos treinados e testados, foram selecionados os top 5 algoritmos com melhores desempenho preditivo, nomeadamente: Random Forest, Histogradient Boosting, LGBM, Bagging, e K vizinhos mais próximos (KNN). A tabela X mostra os resultados do desempenho dos top 5 modelos de machine learning, após terem os seus hiperparâmetros optimizados.

Tabela 1:Resultados para o modelo nos dados de teste

Modelo	R2	MSE	RMSE	MAE
Random Forest	0.8752	0.6611	0.8130	0.6079
HGB	0.8222	0.6816	0.8256	0.6235
LGBM	0.8220	0.6889	0.8300	0.6259
Bagging	0.8638	0.7133	0.8445	0.6137
KNN	0.8222	0.7106	0.8429	0.6318

Nota: RF: random forest; HGB:HistGradientBoosting; MSE: Mean Squared Error; RMSE: Root Mean Squared Error; MAE: mean absolute error

A Figura 3 mostra as curvas dos valores experimentais e preditos (pICC50) da bioatividade dos compostos

contra a proteína CCR-5 para o tratamento potencial de HIV. Podemos observar que os valores de R2 variaram entre 0,82-0,87 mostrando que os cinco modelos de machine learning tiveram uma boa performance na predição da bioatividade contra HIV a partir da estrutura química das moléculas analisadas.

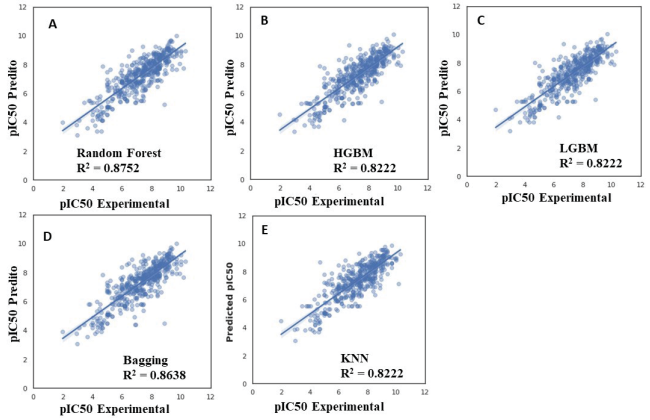

Figura 3: Curva dos valores experimentais (dados de treinamento e de teste) e preditos (dados de validação) da bioatividade dos compostos contra a proteína CCR-5 para o trtamento de HIV.

- **Interpretação dos modelos de machine learning via SHAP values**

A interpretabilidade dos 5 melhores modelos de machine learning (Random Forest, HistGradientBoosting, LGBM, Bagging e KNN) na análise da QSAR para descoberta de novos fármacos para tratamento de HIV foi feita utilizando a abordagem SHAP values [42–44]. Nesta análise o foco, foi a identificação dos descritores moleculares (features) mais importantes no aumento (ou diminuição) da bioatividade dos compostos inibidores da proteína CCR-5 para o tratamento de HIV. A Tabela 2, mostra informações do custo computacional utilizado para o cálculo das permutações feitas pelo Explainer, que é o algoritmo que ajuda a determinar os descritores (features) importantes, que são responsáveis pela bioatividade contra HIV. O KNN foi o algoritmo com maior custo computacional (15 min de execução).

Tabela 2:Desempenho dos diferentes modelos de machine learning no calculo das permutações do explainer

Modelo	N iterações	T gasto	Vel.
RF	470 it	04min:34s	1.64 it/s
HGB	470 it	05min:55s	1.30 it/s
LGBM	470 it	03min:10s	2.33 it/s
Bagging	470 it	06min:59s	1.09 it/s
KNN	470 it	15min:45s	2.04 s/it

Nota: RF: random forest; HGB:HistGradientBoosting; MSE: Mean Squared Error; RMSE: Root Mean Squared Error; MAE: mean absolute error

Na Figura 4, são mostrados os gráficos dos descritores importantes responsaveis pela bioatividade contra HIV. Nos cinco modelos de machine learning as seguintes features mostraram ser importantes na atividade biologica contra HIV: PubchemFP338, PubchemFP451, PubchemFP566, PubchemFP685, PubchemFP335 e PubchemFP13. Os códigos de Smart Pattern e a descrição da substrutrutura são apresentados na tabela 3.

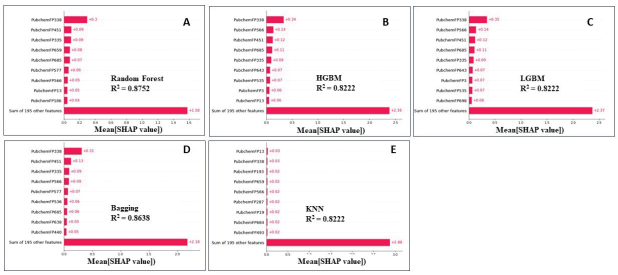

Figura 4: Features importantes dos Top 5 melhores modelos de machine learning na predição da bioatividade inibitória dos compostos contra a proteína CCR-5 para o tratamento de HIV. As features foram determinadas usando a abordagem SHAP values. Podemos observar que as features PubchemFP338, PubchemFP451, PubchemFP566, PubchemFP685, PubchemFP335 e PubchemFP13 estam presentes como importantes em todos cinco modelos de machine learning.

Tabela 3: Resumo dos descritores mais importantes na predicão da atividade biologica dos compostos contra HIV. Todos os descritores listados estavam presentes em todos cinco modelos de machine learning (RF, HGBM, LGBM, KNN). Nesta tabela são mostradas os SMARTS patterns correspondentes e a descrição da sua substrutura.

FSM	Substructure	description
PbFP338	C(C)(C)(H)(N)	N-Isopropilamina
PbFP451	C(-N)(=O)	Formamide
PbFP566	O-C-C-N	Hidroxi amino etano
PbFP685	O=C-C-C-C-N	–
PbFP335	C(C)(C)(C)(H)	n - Butano
PbFP13	32 C	32 átomos carbono

Nota: FSM: Feature Smart Pattern; Pb: Pubchem

A Figura 5, mostra o comportamento dessas features no aumento ou na diminuição da bioatividade (pIC50) para o tratamento de HIV. Desta figura podemos observar a presença do descritor (feature). Nos cinco modelos de machine learning, o descritor PubchemFP338 foi a primeira feature mais importante na predição de bioatividade contra HIV. A presença da feature PubchemFP338 nos compostos aumentou a bioatividade (pIC50) dos mesmos contra a proteína CCR-5 para o tratamento de HIV. Na literatura cientifica, PubchemFP338 representa o grupo funcional amino ligado ao grupo isopropilico (N-isopropilil amina) [39, 45, 46]. Assim, compostos químicos cuja estrutura química apresentam o grupo funcional, em sua estrutura, aumentam a atividade biológica inibitória contra a proteína CCR-5 para o tratamento de HIV. Existem estudos de QSAR também mostrando que o a substrutura N-isopropilil (PubchemFP338) é importante na inibitória da proteína meprin , que é proteína envolvida no desenvolvimento de vários canceres, hyperkeratosis, neurodegenerative disorders, hyperkeratosis, and inflammatory conditions que incluem fisose [45]. Outras features que também impactaram na atividade biológica foi o descritor PubchemFP335, PubchemFP13 e PubchemFP566 (figura 5). A feature PubchemFP335 corresponde a um hidrocarboneto alifático chamado n-butano (Ch3-CH2-CH2-CH3) [47], ao passo que o descritor PubchemFP13 representa número de átomos de carbono maiores ou igual a 32 [48]. E por fim, PubchemFP566 corresponde ao grupo metil-amino etano (OH-CH2-CH2-NH2). A presença dessas três features na estrutura química dos compostos, aumentou, a bioatividade desses compostos na inibição da proteína CCRR-5 para o tratamento do vírus HIV, e vice-versa.

Outra feature mais importante para predição da bioatividade dos compostos contra HIV em nosso estudo é PubchemFP451. PubchemFP451 representa o grupo funcional formamida C(-N)(=O) [49, 50]. . Em nossos todos modelos os cinco modelos machine learning mostraram que a presença da formamida na estrutura química dos compostos químicos, diminuiu a bioatividade desses compostos, por outro lado, a remoção do grupo formamida aumenta a bioatividade dos compostos (figure 5). Assim, uma das estratégias de otimização molecular que poderia ser adotada com vista a potencializar o a atividade biológica dos compostos, seria a remoção do grupo formamida nas estruturas dos compostos candidatos a fármacos inibidores de CCR-5 para o tratamento de HIV [51]. Por fim, a feature PubchemFP566, Outras features que que também

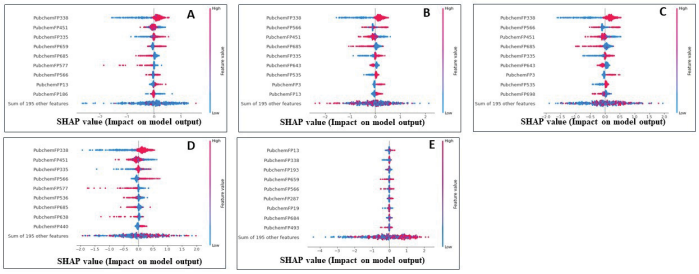

Figura 5: Features importantes dos Top 5 melhores modelos de machine learning na predição da bioatividade inibitória dos compostos contra a proteína CCR-5 para o tratamento de HIV. As features foram determinadas usando a abordagem SHAP values. A presença dos seguintes grupos funcionais na esturutura quimica dos compostos aumentou a bioatividade inibitória contra proteína CCR-5 para o tratamento de HIV contra o HIV: . N-Isopropilamina (feature PubchemFP338), hidroxi-amino-etano (feature PubchemFP566). n-butil (PubchemFP335) e a existência na estrutura química dos compostos número de átomos de carbono maior ou igual 32. Em contranste, a presença do grupo funcional formamida (feature PubchemFP451) nos compostos avaliados, reduziu significativamente a bioatividade.

impactaram na atividade biologica foi o descritor PubchemFP566 que corresponde ao grupo metil-amino etano (OH-CH2-CH2-NH2). A presença e ausência desses grupos funcional e diminui a bioatividade dos compostos contra HIV, respectivamente.

4. Conclusão

Neste estudo, usando dados experimentais de vida real, diversos modelos de machine learning baseados em abordagem QSAR foram treinados, testados e validados visando predizer bioatividade inibitória dos compostos contra a proteína CCR-5 para o tratamento potencial de HIV. Os algoritmo que tiveram maior performance preditiva foram Random Forest, HistGradientBoosting, LGBM, Bagging e KNN, cujos valores de capacidade preditiva (R2) variando entre 82

5. Referências

[1] WHO. HIV data and statistics. https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics.

[2] Lusso P. HIV and the chemokine system: 10 years later. EMBO J 2006; 25: 447–456.

[3] Liu R, Paxton WA, Choe S, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of

some multiply-exposed individuals to HIV-1 infection. Cell 1996; 86: 367–377.

[4] Gulick RM, Lalezari J, Goodrich J, et al. Maraviroc for previously treated patients with R5 HIV-1 infection. New England Journal of Medicine 2008; 359: 1429–1441.

[5] Lederman MM, Penn-Nicholson A, Cho M, et al. Biology of CCR5 and Its Role in HIV Infection and Treatment. JAMA 2006; 296: 815–826.

[6] Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018; 15: 20170387.

[7] Schneider P, Walters WP, Plowright AT, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 2020; 19: 353–364.

[8] Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem 2017; 38: 1291–1307.

[9] Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:151002855.

[10] Witten IH, Frank E, Hall MA, et al. Practical machine learning tools and techniques. Data Mining Fourth Edition, Elsevier Publishers.

[11] Ragoza M, Hochuli J, Idrobo E, et al. Protein–Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model 2017; 57: 942–957.

[12] Zhang L, Tan J, Han D, et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today 2017; 22: 1680–1685.

[13] Lo Y-C, Rensi SE, Torng W, et al. Machine learning in chemoinformatics and drug discovery. Drug Discov Today 2018; 23: 1538–1546.

[14] Cheng F, Li W, Zhou Y, et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties.

[15] Ma J, Sheridan RP, Liaw A, et al. Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inf Model 2015; 55: 263–274.

[16] Tetko I V, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci 1995; 35: 826–833.

[17] Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 2010; 50: 1189.

[18] Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018; 9: 513–530.

[19] Akinleye A, Chen Y, Mukhi N, et al. Ibrutinib and novel BTK inhibitors in clinical development. J Hematol Oncol 2013; 6: 59.

[20] Simeon S, Anuwongcharoen N, Shoombuatong W, et al. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. PeerJ 2016; 4: e2322.

[21] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011; 32: 1466–1474.

[22] Lipinski CA. Lead-and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 2004; 1: 337–341.

[23] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinformatics; 2015.

[24] Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal 2009; 53: 3735–3745.

[25] May RJ, Maier HR, Dandy GC. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Networks 2010; 23: 283–294.

[26] Wei W, Jiang J, Liang H, et al. Application of a combined model with autoregressive integrated moving average (ARIMA) and generalized regression neural network (GRNN) in forecasting hepatitis incidence in Heng County, China. PLoS One 2016; 11: e0156768.

[27] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 2021; 7: e623.

[28] Futagami K, Fukazawa Y, Kapoor N, et al. Pairwise acquisition prediction with SHAP value interpretation. The Journal of Finance and Data Science 2021; 7: 22–44.

[29] Lipinski CA. Lead- and drug-like compounds: The rule-of-five revolution. Drug Discovery Today: Technologies 2004; 1: 337–341.

[30] Varma MVS, Obach RS, Rotter C, et al. Physicochemical space for optimum oral bioavailability: Contribution of human intestinal absorption and first-pass elimination. J Med Chem 2010; 53: 1098–1108.

[31] Sun H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. J Chem Inf Comput Sci 2004; 44: 748–757.

[32] Sandri G, Rossi S, Bonferoni MC, et al. Buccal penetration enhancement properties of N-trimethyl chitosan: Influence of quaternization degree on ab-

sorption of a high molecular weight molecule. Int J Pharm 2005; 297: 146–155.

[33] Chae SY, Jang MK, Nah JW. Influence of molecular weight on oral absorption of water soluble chitosans. Journal of Controlled Release 2005; 102: 383–394.

[34] Rajnikant V, Dinesh J, Bhavnaish C. Biological-activity predictions and hydrogen-bonding analysis of estrane derivatives of steroids. Journal of Chemical Crystallography 2008; 38: 567–576.

[35] Erickson JA, Mcloughlin JI, Lindbergh Boulevard N, et al. Hydrogen Bond Donor Properties of the Difluoromethyl Group, https://pubs.acs.org/sharingguidelines (1995).

[36] Simeon S, Anuwongcharoen N, Shoombuatong W, et al. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. PeerJ; 2016. Epub ahead of print 2016. DOI: 10.7717/PEERJ.2322.

[37] Chtita S, Ghamali M, Ousaa A, et al. QSAR study of anti-Human African Trypanosomiasis activity for 2-phenylimidazopyridines derivatives using DFT and Lipinski's descriptors. DOI: 10.1016/j.heliyon.2019.

[38] Paliwal S, Seth D, Yadav D, et al. Development of a robust QSAR model to predict the affinity of pyrrolidine analogs for dipeptidyl peptidase IV (DPP-IV). J Enzyme Inhib Med Chem 2011; 26: 129–140.

[39] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011; 32: 1466–1474.

[40] Senawi A, Wei HL, Billings SA. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. Pattern Recognit 2017; 67: 47–61.

[41] Lindner T, Puck J, Verbeke A. Beyond addressing multicollinearity: Robust quantitative analysis and machine learning in international business research. Journal of International Business Studies 2022; 53: 1307–1314.

[42] Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018; 2: 749–760.

[43] Wojtuch A, Jankowski R, Podlewska S. How can SHAP values help to shape metabolic stability of chemical compounds? J Cheminform; 13. Epub ahead of print 1 December 2021. DOI: 10.1186/s13321-021-00542-y.

[44] Ding Y, Chen M, Guo C, et al. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. J Mol Liq; 326. Epub ahead of print 15 March 2021. DOI: 10.1016/j.molliq.2020.115212.

[45] Banerjee S, Baidya SK, Ghosh B, et al. Quantitative structural assessments of potential meprin inhibitors by non-linear QSAR approaches and validation by binding mode of interaction analysis. New Journal of Chemistry 2023; 47: 7051–7069.

[46] Saini R, Fatima S, Agarwal SM. TMLRpred: A machine learning classification model to distinguish reversible EGFR double mutant inhibitors. Chem Biol Drug Des 2020; 96: 921–930.

[47] Malik AA, Phanus-umporn C, Schaduangrat N, et al. HCVpred: A web server for predicting the bioactivity of hepatitis C virus NS5B inhibitors. J Comput Chem 2020; 41: 1820–1834.

[48] Yang M, Chen J, Shi X, et al. Development of in silico models for predicting p-glycoprotein inhibitors based on a two-step approach for feature selection and its application to Chinese herbal medicine screening. Mol Pharm 2015; 12: 3691–3713.

[49] Schaduangrat N, Malik AA, Nantasenamat C. ERpred: a web server for the prediction of subtype-specific estrogen receptor antagonists. PeerJ; 9. Epub ahead of print 1 July 2021. DOI: 10.7717/peerj.11716.

[50] Fernandes S, Chong JJH, Paige SL, et al. Comparison of human embryonic stem cell-derived cardiomyocytes, cardiovascular progenitors, and bone marrow mononuclear cells for cardiac repair. Stem Cell Reports 2015; 5: 753–762.

[51] Nicolaou CA, Brown N, Pattichis CS. Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Devel 2007; 10: 316–324.