

QSAR DE INIBIDORES DA JANUS QUINASE 2: UMA ABORDAGEM COM INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Matheus Paulo Bastos, Ana Júlia Rodrigues Lacerda, Sheida Mehrpour, Othon Souto Campos, Heberth de Paula.

Universidade Federal do Espírito Santo, Alto Universitário, s/n, Guararema - 29500-000 - Alegre-ES, Brasil, mathpbastos@gmail.com, anajulia.r09@gmail.com, sheyda.mehrpour@gmail.com, othon.campos@ufes.br, heberth.paula@ufes.br.

Resumo

A inibição da *Janus quinase 2* (JAK2) é uma estratégia terapêutica promissora para o tratamento de diversas neoplasias hematológicas. Para otimizar a descoberta de novos compostos, este trabalho descreveu a construção de modelos de *Quantitative Structure-Activity Relationship* (QSAR). Um conjunto de dados contendo 7.233 compostos com valores de bioatividade (IC₅₀) foi extraído da base de dados BindingDB. Após um processo de tratamento e conversão de dados para a escala logarítmica (pIC₅₀), o conjunto foi dividido em dois subconjuntos de treinamento (80%) e teste (20%) pelo algoritmo Kennard-Stone, garantindo a representatividade química em ambos. A partir do conjunto de treinamento, foi construído um modelo de aprendizado de máquina baseado no algoritmo XGBoost e métodos de inteligência artificial explicáveis. O poder preditivo do modelo foi avaliado externamente com os dados de teste, que resultou em um coeficiente de determinação (R²) de 0,646 e um erro médio absoluto (MAE) de 0,492. Esse estudo demonstra o potencial de métodos de aprendizado de máquina e na descoberta de descritores que atuam na inibição da JAK2.

Palavras-chave: Machine Learning. QSAR. JAK2. XGBoost. Inteligência Artificial Explicável.

Área do Conhecimento: Ciências da Saúde. Farmácia.

Introdução

As quinases JAK são uma família de tirosino quinases não receptoras intracelulares que incluem JAK1, JAK2, JAK3 e TYK2. Estas quinases são responsáveis pela transdução de sinais mediados por citocinas, por meio da via JAK-STAT (SINGH; NAVEENA; KARTHIKEYAN, 2014). Assim, exercem um papel crucial na proliferação, sobrevivência e diferenciação de diferentes tipos celulares (CONSTANTINESCU; GIRARDOT; PECQUET, 2008). Assim, altos níveis de citocinas e mediadores de sinalização são encontrados em diversas neoplasias hematológicas (VERSTOVSEK, 2009). Diante disso, a descoberta de potenciais inibidores de JAK2 pode ser um grande avanço para a indústria farmacêutica no tratamento de diversas doenças.

Para avaliar potenciais novos fármacos são utilizadas técnicas *in silico*, como o *Quantitative Structure-Activity Relationship* (QSAR), no qual inclui aspectos químicos, físicos, biológicos e toxicológicos (GINI, 2024). Além disso, permite a exploração de características químicas codificadas por meio de descritores moleculares, com o intuito de estabelecer uma correlação matemática entre uma resposta química e atributos químicos quantitativos que definem as características da molécula (ROY; KAR; DAS, 2015). Para isso, bancos de dados químicos contendo milhões de compostos testados, como o BindingDB e PubCHEM estão cada vez maiores (TROPSHA, 2010).

Metodologia

As estruturas moleculares e seus respectivos dados de bioatividade foram obtidos da base de dados BindingDB (CHEN; LIU; GILSON, 2001). Um filtro foi aplicado, a fim de remover redundância e manter apenas moléculas que apresentassem valores de atividade biológica bem definidos, em que valores de IC₅₀ estivessem expressos em nanomolar (nM). Os códigos SMILES das moléculas foram padronizados para sua forma canônica por meio da biblioteca RDKit (<http://www.rdkit.org>). O software PaDEL-Descriptor (YAP, 2011) foi empregado para o cálculo dos descritores moleculares 1D, 2D e 881 fingerprints do tipo PubChem.

Para assegurar a qualidade dos dados de bioatividade e remover compostos conhecidos por causarem interferência em ensaios bioquímicos, foi aplicado um filtro para identificar e remover as

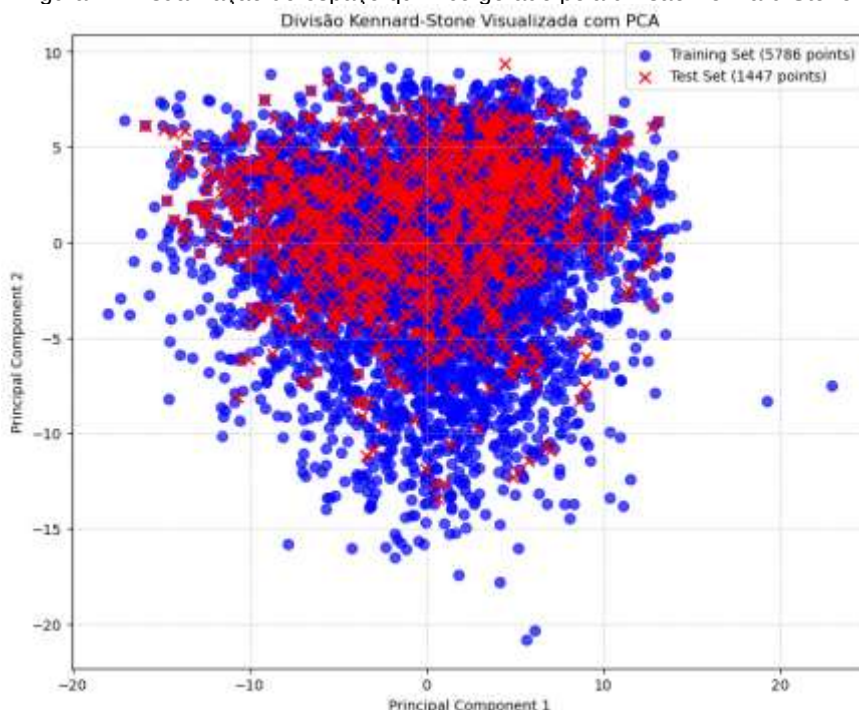
moléculas que apresentassem subestruturas do tipo PAINS (*Pan-Assay Interference Compounds*), conforme proposto por Baell & Holloway (2010). A fim de avaliar o potencial farmacocinético e a propriedade *drug-likeness* dos compostos, as moléculas filtradas foram submetidas ao filtro de Lipinski com o intuito de selecionar as moléculas que não violassem uma das seguintes regras: peso molecular (MW) inferior ou igual 500Da, coeficiente de partição octanol-água (AlogP) não superior a 5, número de doadores de ligações de hidrogênio (nHBDon) inferior ou igual a 5; e o número de aceptores de ligações de hidrogênio (nHBAcc) inferior ou igual a 10 (LIPINSKI, 1997).

Na análise exploratória e manipulação dos dados, foram utilizadas as bibliotecas pandas, numpy, matplotlib. A ferramenta Scikit-learn (PEDREGOSA *et al.*, 2011) foi empregada a fim de aplicar técnicas estatísticas descritivas e transformação de características. As instâncias com valores nulos e as características com correlação superior a 0,7 foram removidas. O cálculo de seus respectivos valores de pIC50 foi realizado para as moléculas selecionadas, através da equação:

$$pIC_{50} = -\log_{10}([IC_{50}] \times 10^{(-9)})$$

O algoritmo Kennard Stone (KENNARD; STONE, 1969) foi aplicado para selecionar um subconjunto de compostos para o conjunto de teste, a fim de obter uma cobertura uniforme e representativa do espaço químico definido pelos descritores moleculares. Com isso, duas amostras distantes, de acordo com a distância geométrica, estabeleceram os limites. Em seguida, a amostra seguinte foi inserida ao subconjunto para selecionar aquelas que obtiveram maior separação. Este processo foi repetido até que o número estipulado de amostras fosse atingido. A partir disso, uma seleção uniformemente distribuída e espaçada sobre o conjunto de dados foi realizada, incluindo instâncias que definem as fronteiras do conjunto de dados (Figura 1) (ANDRADA, 2017).

Figura 1 - Visualização do espaço químico gerado pela divisão Kennard Stone.



Fonte: O autor (2025).

Para construção do modelo, foi utilizado o algoritmo XGBoost (*eXtreme Gradient Boosting*) (CHEN; GUESTRIN, 2016), um algoritmo de aprendizado de máquina do tipo *ensemble*, baseado em árvores de decisão. A escolha deste algoritmo é justificada por sua comprovada capacidade de capturar relações complexas e não-lineares entre os descritores moleculares e suas respectivas atividades biológicas. Para identificar a configuração ótima e maximizar o poder preditivo, foi empregado o *framework* de otimização de hiperparâmetros Optuna (AKIBA *et al.*, 2019), no qual foi empregado um algoritmo *Tree-structured Parzen Estimator* (TPE), um tipo de otimização Bayesiana, a fim de modelar

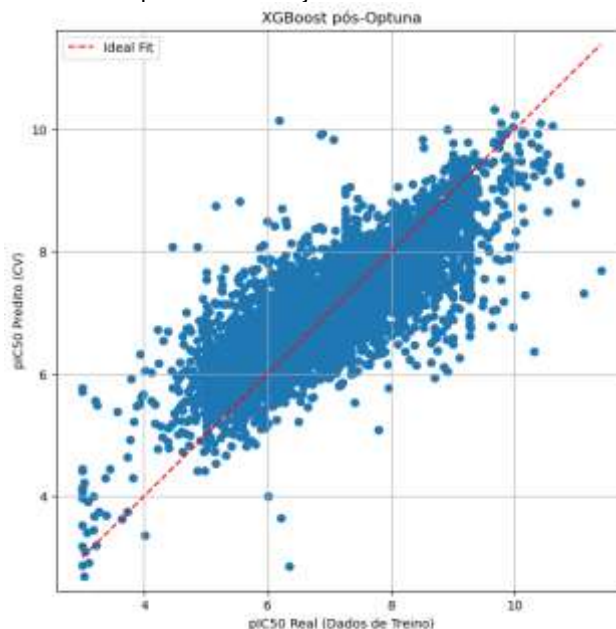
probabilisticamente o espaço de busca e guiar a amostragem de novas configurações de forma inteligente, equilibrando eficientemente a exploração e a exploração do espaço de hiperparâmetros. O modelo final, após a otimização com Optuna, foi avaliado utilizando as métricas de Coeficiente de Determinação (R^2), Erro Médio Absoluto (MAE) e Raiz do Erro Quadrático Médio (RMSE).

Para a interpretabilidade do modelo, considerado "caixa-preta" (*black box*), foi empregada a metodologia SHAP (*Shapley Additive Explanations*), uma abordagem unificada que atribuiu para cada descritor molecular (*feature*) um valor de importância em cada predição individual (LUNDBERG; LEE, 2017). Assim, foi quantificada a contribuição marginal de um descritor para afastar a predição da média de todas as predições do modelo e, por fim, estes valores foram utilizados para gerar gráficos de resumo que visualizam a relação entre o valor de um descritor e seu impacto no modelo.

Resultados

A avaliação de dois modelos XGBoost distintos evidenciou o impacto positivo da otimização de hiperparâmetros na performance preditiva. O modelo otimizado, resultado da aplicação da metodologia Optuna, demonstrou consistentemente um desempenho superior em relação ao modelo de base (*baseline*), cujos hiperparâmetros foram estabelecidos manualmente. A correlação entre os valores observados e preditos para o modelo otimizado é apresentada na Figura 2, enquanto a análise comparativa das métricas estatísticas foram detalhadas na Tabela 1.

Figura 2 - Diagrama de dispersão dos valores reais de pIC50 (eixo x) e os valores preditos pelo modelo XGBoost após otimização de hiperparâmetros com a metodologia Optuna (eixo y). A referência 'Ideal Fit' representada pela linha tracejada vermelha.



Fonte: O autor (2025).

Tabela 1 - Métricas estatísticas do modelo XGBoost antes do processo de otimização.

Métricas	Treino	Teste
R^2	0.842 (+/- 0.002)	0.601 (+/- 0.018)
MAE	0.356 (+/- 0.002)	0.565 (+/- 0.011)
RMSE	0.469 (+/- 0.003)	0.746 (+/- 0.020)

Fonte: O autor.

Os resultados da validação externa do modelo XGBoost otimizado através de um conjunto de teste exclusivo, a fim de garantir que as moléculas analisadas não tivessem qualquer exposição prévia ao treinamento, foram apresentados na Tabela 2.

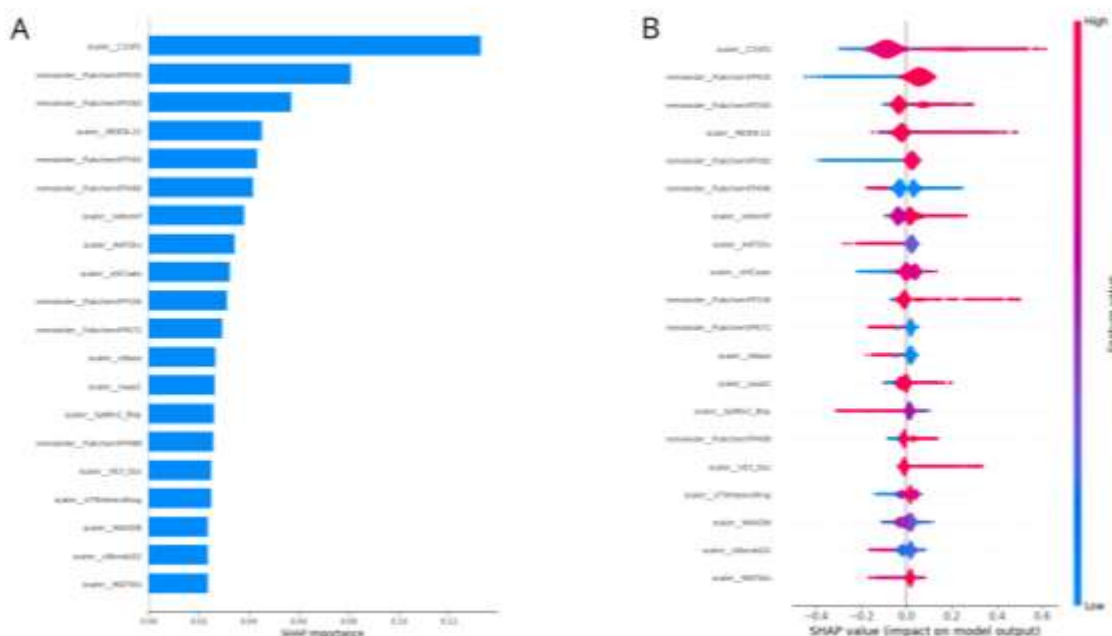
Tabela 2 - Métricas estatísticas do modelo XGBoost após o processo de otimização.

Métricas	Treino (CV)	Teste (CV)	Validação Externa
R ²	0.936 (+/- 0.001)	0.644 (+/- 0.025)	0.646
MAE	0.221 (+/- 0.001)	0.522 (+/- 0.015)	0.672
RMSE	0.298 (+/- 0.002)	0.746 (+/- 0.027)	0.492

Fonte: O autor.

Os gráficos de resumo SHAP apresentam os 20 descritores ou *fingerprints* de forma ordenada, sendo apresentados do maior para o menor impacto, além da direção e magnitude de seus efeitos (Figura 3). Na análise, o descritor C1SP2 obteve destaque com uma correlação positiva direta com a atividade e a presença das subestruturas codificadas pelos *fingerprints* PubChemFP435 e PubChemFP260 foi avaliada como um forte preditor de alta atividade.

Figura 3 - 20 principais descritores apontados pelo SHAP. (A) Ordem de importância bem como sua contribuição com o modelo. (B) Magnitude da contribuição dos principais descritores.



Fonte: O autor.

Discussão

A interpretabilidade do modelo QSAR, avaliada pela metodologia SHAP, forneceu *insights* valiosos sobre as características moleculares que governam a inibição da JAK2. A análise foi liderada pelo descritor C1SP2, que quantifica a presença de carbonos sp² ligados a um único átomo de carbono, uma característica comum em grupos funcionais aceptores de ligação de hidrogênio, como as carbonilas (C=O). Os gráficos SHAP demonstraram uma relação direta e robusta: valores elevados de C1SP2 consistentemente aumentam a potência predita, enquanto sua ausência está associada a uma menor atividade. Esta observação alinha-se ao mecanismo de ação estabelecido para inibidores de

quinase proposto por Raivola e colaboradores (2021), no qual os inibidores clinicamente relevantes atuam competindo com o ATP no domínio quinase da JAK.

O princípio fundamental por trás dessa inibição é a formação de interações-chave, como ligações de hidrogênio com a região da dobradiça (*hinge region*), que são essenciais para a regulação da atividade da enzima (BABON *et al.*, 2014). Para além das propriedades gerais, a análise SHAP também destacou a relevância de subestruturas moleculares específicas (Tabela 3), representadas por descritores do tipo *fingerprint*. A importância do PubChemFP260 indicou que a presença de múltiplos anéis heteroaromáticos é uma característica favorável, que reforçam a estrutura de inibidores complexos e potentes como o Lestaurtinib (HEXNER *et al.*, 2008). Da mesma forma, a relevância do PubChemFP392 reflete o papel estrutural de conectores (*linkers*) do tipo amina secundária, uma característica presente no fármaco aprovado Fedratinib (WERNIG *et al.*, 2008). Finalmente, o fragmento aldimina, identificado pelo PubChemFP435, atua como um aceptor de ligação de hidrogênio adicional. Embora este fragmento específico não seja encontrado nos inibidores mais proeminentes, sua importância preditiva reforça o princípio fundamental de que múltiplas interações via ligação de hidrogênio no sítio ativo são cruciais para a otimização da afinidade e, conseqüentemente, da potência inibitória (BABON *et al.*, 2014). Em conjunto, esses resultados demonstram que o modelo QSAR desenvolvido não é apenas uma ferramenta preditiva, mas também um gerador de hipóteses quimicamente coerentes e alinhadas ao conhecimento farmacológico atual.

Tabela 3 - *Fingerprints* e suas subestruturas.

<i>Fingerprints</i>	Subestrutura
PubchemFP260	≥ 3 anéis heteroaromáticos
PubchemFP392	N(∼C)(∼C)(∼H)
PubchemFP435	C(∼C)(∼H)(=N)

Fonte: PubChem

Conclusão

Este trabalho aplicou com sucesso um modelo de Inteligência Artificial Explicável (XAI) no campo da bioinformática para identificar as características moleculares determinantes para a atividade de inibidores da JAK2. A análise SHAP do modelo QSAR revelou que a potência inibitória é governada por uma combinação de propriedades físico-químicas, como a presença de aceptores de ligação de hidrogênio, e pela presença de subestruturas específicas, como *scaffolds* heteroaromáticos validados na literatura. Portanto, não apenas gerou um modelo preditivo robusto, mas também demonstrou o potencial da XAI como uma ferramenta poderosa para extrair conhecimento químico acionável, acelerando o planejamento de novos candidatos terapêuticos.

Referências

- AKIBA, T. *et al.* Optuna: A next-generation hyperparameter optimization framework. **Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining**. 2019. p. 2623-2631.
- ANDRADA, M.F *et al.* Impact assessment of the rational selection of training and test sets on the predictive ability of QSAR models. **SAR and QSAR in Environmental Research**, v. 28, n. 12, p. 1011-1023, 2017.
- BABON, J.J. *et al.* The molecular regulation of Janus kinase (JAK) activation. **Biochemical Journal**, v. 462, n. 1, p. 1-13, 2014.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. 2016. p. 785-794.

CHEN, X.; LIU, M.; GILSON, M.K. BindingDB: a web-accessible molecular recognition database. **Combinatorial chemistry & high throughput screening**, v. 4, n. 8, p. 719-725, 2001.

CONSTANTINESCU, S.N.; GIRARDOT, M.; PECQUET, C. Mining for JAK-STAT mutations in cancer. **Trends in biochemical sciences**, v. 33, n. 3, p. 122-131, 2008.

GINI, G.C. QSAR: Using the Past to Study the Present. **Computational Toxicology: Methods and Protocols**, p. 3-39, 2024.

HEXNER, E.O. *et al.* Lestaurtinib (CEP701) is a JAK2 inhibitor that suppresses JAK2/STAT5 signaling and the proliferation of primary erythroid cells from patients with myeloproliferative disorders. **Blood, The Journal of the American Society of Hematology**, v. 111, n. 12, p. 5663-5671, 2008.

KENNARD, R.W.; STONE, L.A. Computer aided design of experiments. **Technometrics**, v. 11, n. 1, p. 137-148, 1969.

LANDRUM, G. *et al.* Rdkit: Open-Source Chmeinformatics Software, 2016. Zenodo. Disponível em: <http://www.rdkit.org>. Acesso em: 30 ago. 2025.

LIPINSKI, C.A. *et al.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **Advanced drug delivery reviews**, v. 23, n. 1-3, p. 3-25, 1997.

LUNDBERG, S.M.; LEE, S. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

PEDREGOSA, F. *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

RAIVOLA, J. *et al.* Janus kinases in leukemia. **Cancers**, v. 13, n. 4, p. 800, 2021.

ROY, K.; KAR, S.; DAS, R.N. A primer on QSAR/QSPR modeling: fundamental concepts. **Springer**, 2015.

SINGH, K.; NAVEENA, Q.; KARTHIKEYAN, M. Jak2 inhibitor – a jackpot for pharmaceutical industries: a comprehensive computational method in the discovery of new potent Jak2 inhibitors. **Molecular BioSystems**, v. 10, n. 8, p. 2146-2159, 2014. DOI: <https://doi.org/10.1039/c4mb00071d>

TROPSHA, A. Best practices for QSAR model development, validation, and exploitation. **Molecular informatics**, v. 29, n. 6- 7, p. 476-488, 2010.

VERSTOVSEK, S. Therapeutic potential of JAK2 inhibitors. **ASH Education Program Book**, v. 2009, n. 1, p. 636-642, 2009.

WERNIG, G. *et al.* Efficacy of TG101348, a selective JAK2 inhibitor, in treatment of a murine model of JAK2V617F-induced polycythemia vera. **Cancer cell**, v. 13, n. 4, p. 311-320, 2008.

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, 2011.