



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM PRODUTOS
NATURAIS E SINTÉTICOS BIOATIVOS



LUIS FELIPE DE MORAIS MELO

**AmIActive (AIA): Uma ferramenta web preditiva de larga
escala baseada em QSAR para identificação de alvos e
polifarmacologia**

JOÃO PESSOA

2026

LUIS FELIPE DE MORAIS MELO

**AmIActive (AIA): Uma ferramenta web preditiva de larga
escala baseada em QSAR para identificação de alvos e
polifarmacologia**

Dissertação de mestrado do Programa de
Pós-Graduação em Produtos Naturais e
Sintéticos Bioativos da Universidade
Federal da Paraíba, apresentada como
requisito para obtenção do título de
Mestre em Produtos Naturais e Sintéticos
Bioativos. Área de concentração:
Farmacoquímica

Orientador: Prof. Dr. Marcus Tullius Scotti

JOÃO PESSOA

2026

Catálogo na publicação
Seção de Catalogação e Classificação

M528a Melo, Luis Felipe de Moraes.

AmIActive (AIA) : uma ferramenta web preditiva de
larga escala baseada em QSAR para identificação de
alvos e polifarmacologia / Luis Felipe de Moraes Melo.
- João Pessoa, 2026.
68 f. : il.

Orientação: Marcus Tullius Scotti.
Dissertação (Mestrado) - UFPB/CCS.

1. Triagem reversa. 2. Triagem de alvo. 3.
Aprendizado de máquina. 4. Quimioinformática. 5.
Triagem virtual baseada em ligante. I. Scotti, Marcus
Tullius. II. Título.

UFPB/BC

CDU 004.89:615(043)



Ata da 526ª (quingentésima vigésima sexta) Dissertação de Mestrado do(a) aluno(a) do Programa de Pós-Graduação em Produtos Naturais e Sintéticos Bioativos **Luis Felipe de Moraes Melo**, candidato(a) ao Título de “Mestre” em Produtos Naturais e Sintéticos Bioativos na área de concentração Farmacoquímica.

Às quinze horas (15h00) do dia vinte de fevereiro do ano de dois mil e vinte e seis (20/02/2026), em ambiente virtual de videoconferência através do aplicativo Google Meet, link: <https://meet.google.com/bdr-zjir-gqz>, reuniram-se em caráter de Solenidade Pública os membros da Comissão designada para examinar o(a) discente **Luis Felipe de Moraes Melo**, candidato(a) ao Título de “MESTRE” em Produtos Naturais e Sintéticos Bioativos na área de concentração Farmacoquímica. Foram componentes da Comissão Examinadora os pesquisadores Eugene Muratov, Ph.D em Química Orgânica, Francisco Jaime Bezerra Mendonça Junior, Ph.D em Ciências Biológicas e e Marcus Tullius Scotti, Ph.D em Química Orgânica. Sendo o primeiro, integrante do corpo docente da Universidade da Carolina do Norte, o segundo, integrante do corpo docente da Universidade Estadual da Paraíba e o último, integrante do corpo docente da Universidade Federal da Paraíba. Dando início aos trabalhos, o(a) Presidente da Comissão, professor(a) Marcus Tullius Scotti, após declarar os objetivos da reunião, apresentou o(a) candidato(a) **Luis Felipe de Moraes Melo**, a quem concedeu a palavra para que dissertasse oral e sucintamente sobre o tema apresentado e intitulado “AmIActive (AIA): Uma ferramenta web preditiva de larga escala baseada em QSAR para identificação de alvos e polifarmacologia”. Após discorrer sobre o referido tema durante cerca de quarenta minutos, o(a) candidato(a) foi arguido(a) pelos examinadores na forma regimental. Em seguida, passou a comissão, em caráter secreto, a proceder à avaliação e julgamento do trabalho, concluindo por atribuir-lhe o conceito APROVADO(A). Em face da aprovação, declarou o(a) Presidente achar-se o(a) examinado(a) **Luis Felipe de Moraes Melo** legalmente habilitado(a) a receber o Título de “MESTRE” em Produtos Naturais e Sintéticos Bioativos, na área de concentração Farmacoquímica, cabendo a Universidade Federal da Paraíba, providências, como de direito, a expedição do Diploma que o(a) mesmo(a) faz jus. Nada mais havendo a tratar, foi lavrada a presente ata que é abaixo assinada pelos membros da Comissão e pelo(a) discente.

Prof. Dr. Marcus Tullius Scotti (Orientador)



Documento assinado digitalmente
MARCUS TULLIUS SCOTTI
Data: 23/02/2026 21:54:07-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Eugene Muratov (Examinador)

Eugene Muratov

Prof. Dr. Francisco Jaime Bezerra Mendonça Junior (Examinador)



Documento assinado digitalmente
FRANCISCO JAIME BEZERRA MENDONÇA JUNIOR
Data: 23/02/2026 23:02:53-0300
Verifique em <https://validar.iti.gov.br>

Luis Felipe de Moraes Melo (Discente)



Documento assinado digitalmente
LUIS FELIPE DE MORAIS MELO
Data: 24/02/2026 11:02:54-0300
Verifique em <https://validar.iti.gov.br>

Dedico este trabalho aos meus pais,
cujo esforço e paciência excepcionais
em minha criação foram essenciais
para que eu chegasse até aqui.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pois foram incontáveis as oportunidades, desafios, conquistas, aprendizados e acontecimentos fortuitos durante a minha jornada acadêmica até então.

Agradeço aos meus pais e minha família, pois sem seu apoio direto e indireto eu definitivamente não teria chegado até aqui. Agradeço em especial a minha mãe, que desde do falecimento do meu Pai, teve de ser tanto pai quanto mãe.

Agradeço também a minha namorada Williane cuja presença em minha vida foi absolutamente diferencial para a esta etapa da minha vida. Sem ela, não sei como teria passado pelos inúmeros desafios e momentos de dificuldade a longo da minha trajetória acadêmica.

Agradeço a meus grandes amigos da vida, alguns de mais de dez anos. São eles: Lucas, Henrique, Xande, Igã, Jops, Murilo, Davi, Kayo e Brunão. Muito obrigado pela amizade, pelos momentos e pela solicitude durante todos esses anos todos.

Gostaria de agradecer ao meu orientador, o Professor Marcus Tullius Scotti. Sem sua orientação eu definitivamente não teria conseguido chegar até este ponto e minha vida acadêmica. Desde que bati em sua porta quando ainda era somente um graduando confuso, fui exposto a múltiplas oportunidades e conhecimentos diferentes, em mundo que eu, até então, desconhecia.

Agradeço também a todos os membros do laboratório de quimioinformática da UFPB, Em especial aos meus colegas Uilames, Eduardo e Luciano, pois sem suas contribuições este trabalho não seria possível. Também agradeço aos membros titulares e suplentes da banca avaliadora, o professor Eugene Muratov, professor Francisco Jaime Bezerra Mendonça Júnior, professor Sócrates Golzio dos Santos e o professor Luis Cezar Rodrigues pela disponibilidade de avaliar o meu trabalho. Por fim, deixo meus agradecimentos a UFPB e a PqPNSB na figura todos os seus profissionais e em especial as secretárias do programa, pela paciência e solicitude em atender os discentes.

RESUMO

O planejamento de fármacos auxiliado por computador (*computer-aided drug design*, CADD) tornou-se um pilar da descoberta moderna de medicamentos, com abordagens como métodos baseados em ligantes (por exemplo, modelagem QSAR-*Quantitative Structure Activity Relationships*, análise de farmacóforos) e métodos baseados em estrutura (por exemplo, docking molecular, dinâmica molecular) impulsionando a inovação. As estratégias baseadas em ligantes permanecem essenciais, especialmente diante da crescente disponibilidade de dados de atividade biológica em larga escala. Em razão disto, ao longo dos anos, diversas plataformas preditivas baseadas em QSAR disponíveis na internet, moldaram a área, tais como PASS, PLATO, SwissTargetPrediction, dentre outros. Embora este ecossistema tenha crescido continuamente desde sua concepção, devido à natureza dinâmica das metodologias computacionais associadas e da grande diversidade de possíveis abordagens, sempre houve espaço para criação de novas ferramentas e a otimização das ferramentas existentes. Deste modo, este estudo teve como objetivo criar a primeira ferramenta brasileira deste tipo. Nesse contexto, apresentamos o AmIActive (AIA), uma ferramenta web preditiva em larga escala baseada em QSAR, projetada para facilitar estudos de target fishing e a avaliação polifarmacológica para alvos moleculares e não moleculares. No capítulo 1, apresenta-se uma revisão das metodologias e conceitos teóricos fundamentais para a compreensão técnica e conceitual de um sistema QSAR deste tipo. O capítulo 2 demonstra o desenvolvimento do sistema AIA propriamente dito. O sistema AIA integra 3.239 modelos QSAR estatisticamente validados, cobrindo 2.277 alvos distintos, incluindo proteínas individuais, complexos proteicos, famílias de proteínas, linhagens celulares, organismos e tecidos. Os dados foram extraídos do ChEMBL30 e submetidos a rigorosos processos de limpeza e padronização, com exclusão de conjuntos inadequados e conversão das atividades para escalas logarítmicas negativas (pIC50, pAC50). Os limiares de classificação entre compostos ativos e inativos foram definidos por um algoritmo de busca binária visando ao balanceamento das classes. Os descritores moleculares ECFP4 (1024 bits) foram utilizados como variáveis de entrada, e os modelos foram construídos com o algoritmo *Random Forest*, com otimização bayesiana de hiperparâmetros e validação cruzada aninhada. Apenas modelos com valor $p < 0,05$ e $MCC \geq 0,5$ foram retidos, apresentando desempenho médio robusto (acurácia $\approx 0,82$; F1-score $\approx 0,82$; $MCC \approx 0,65$). Os domínios de aplicabilidade foram definidos por distância euclidiana normalizada, em conformidade com as diretrizes da OECD (*Organisation for Economic Co-operation and Development*) para QSAR. O sistema está disponível gratuitamente em <https://amiactive.ccen.ufpb.br/>, com interface gráfica baseada no MarvinJS, permitindo a entrada de estruturas ou SMILES e a exportação dos resultados em formato CSV, incluindo probabilidades preditas, limiares de decisão, confiabilidade do domínio de aplicabilidade e métricas de validação dos modelos.

Palavras-chave: Triagem de alvo, Triagem reversa, Aprendizado de máquina, Químioinformática, Triagem virtual baseada em ligante.

ABSTRACT

Computer-aided drug design (CADD) has become a cornerstone of modern drug discovery, with approaches such as ligand-based methods (e.g., QSAR-Quantitative Structure Activity Relationships modeling and pharmacophore analysis) and structure-based methods (e.g., molecular docking and molecular dynamics) driving innovation. Ligand-based strategies remain essential, particularly in light of the growing availability of large-scale biological activity data. As a result, over the years, several QSAR-based predictive platforms available on the internet have shaped the field, including PASS, PLATO, and SwissTargetPrediction, among others. Although this ecosystem has continuously expanded since its inception, the dynamic nature of computational methodologies and the wide diversity of possible approaches have consistently left room for the development of new tools and the optimization of existing ones. Accordingly, the aim of this study was to develop the first Brazilian tool of this type. In this context, we present AmIActive (AIA), a large-scale QSAR-based predictive web tool designed to facilitate target fishing studies and polypharmacology assessment for both molecular and non-molecular targets. Chapter 1 presents a review of the fundamental methodologies and theoretical concepts required for the technical and conceptual understanding of a QSAR system of this type, while Chapter 2 describes the development of the AIA system itself. The AIA system integrates 3,239 statistically validated QSAR models covering 2,277 distinct targets, including single proteins, protein complexes, protein families, cell lines, organisms, and tissues. Pharmacological data were extracted from ChEMBL30 and subjected to rigorous cleaning and standardization procedures, with the exclusion of unsuitable datasets and conversion of activity values to negative logarithmic scales (pIC₅₀, pAC₅₀). Classification thresholds separating active and inactive compounds were defined using a binary search algorithm to achieve balanced classes. ECFP4 molecular fingerprints (1024 bits) were used as input features, and models were built using the Random Forest algorithm, with Bayesian hyperparameter optimization and nested cross-validation. Only models with p-values < 0.05 and MCC ≥ 0.5 were retained, yielding robust average performance (accuracy ≈ 0.82, F1-score ≈ 0.82, MCC ≈ 0.65). Applicability domains were defined using a normalized Euclidean distance metric, in accordance with OECD (Organisation for Economic Co-operation and Development) QSAR guidelines. The system is freely available at <https://amiactive.ccen.ufpb.br/> through an intuitive graphical interface based on ChemAxon's MarvinJS, allowing users to draw chemical structures or input SMILES strings and export prediction results in CSV format, including predicted activity probabilities, classification thresholds, applicability domain reliability, and complete model validation metrics.

Keywords: Target screening, Reverse Screening, Machine learning, Chemoinformatics, ligand-based virtual screening.

LISTA DE ABREVIATURAS, SIGLAS E FÓRMULAS

AIA – AmIActive

Bagging – Bootstrap Aggregation (Agregação por amostragem com repetição)

CADD – Computer Aided Drug Design (Design de drogas assistido por computador)

CHEMBL – Chemogenomics database developed at the European Molecular Biology Laboratory

CSV – Comma Separated Values (Valores separados por vírgula)

EC₅₀ – Effective Concentration 50% (Concentração Efetiva 50%)

ECFP – Extended Connectivity Fingerprints

GUSAR – General Unrestricted Strcuture-Activity Relationships

IC₅₀ – Inhibition Concentration 50% (Concentração Inibitória 50%)

K_i – Inhibition Constant (Constante de inibição)

LIME – Local interpretable model-agnositic explanations

MAE – Mean Absolute Error (Erro absoluto médio)

MACCS – Molecular Access System

MCC/CCM – Matthews Correlation Coefficient/Coeficiente de correlação de Matthews

MSE – Mean Squared Error (Erro quadardo médio)

nM – Nano molar

uM – Micro molar

OECD – Organisation for Economic Co-operation and Development (Organização para a Cooperação e Desenvolvimento Econômico)

PASS – Prediction of Activity Spectra for Substances

PLATO – Polypharmacology plataform prediction

QNA – Quantitative Neighborhoods of Atoms

QSAR – Quantitative Structure Activity Relationships (Relações quantitativas estrutura atividade)

RBF – Radial basis function artificial neural networks (Redes neurais de função de base radial)

RF – Random Forest (Floresta aleatória)

RMSE – Root Mean Squared Error (Raiz do erro quadrado médio)

SCR – Self-consistent regression

SHAP – Shapley Additive Explanations

SMILES – Simplified Molecular Input Line Entry System

TPE – Tree-structured parzen estimators

LISTA DE FIGURAS

CAPÍTULO I

Figura 1: Esquematização do funcionamento do aprendizado de máquina.....	17
Figura 2: Exemplos de problemas de regressão e classificação.....	19
Figura 3: Exemplo de sobreajuste.....	23
Figura 4: Esquematização do procedimento de validação cruzada.....	25
Figura 5: Esquematização de uma árvore de decisão.....	25
Figura 6: Esquematização de uma floresta aleatória.....	26
Figura 7: Esquematização de um Fingerprint genérico.....	28
Figura 8: Esquematização do processo de treinamento de modelos QSAR.....	29

CAPÍTULO II

Figure 1: Detailed Schematization of the modeling process of the AIA system.....	37
Figure 2: The AIA system user interface; A) The Marvin JS applet; B) The generate smiles button; C) The smiles text box; D) The predict button; E) The help and models info buttons.....	39
Supplementary figure 1: Comparison of the AIA, PLATO, SwissTargetPrediction and PASS systems performance on four targets.....	45

LISTA DE TABELAS

CAPÍTULO II

Table 2: Average thresholds and validation metrics of the QSAR models grouped by target type.....	39
--	-----------

SUMÁRIO

CAPÍTULO I	14
INTRODUÇÃO.....	15
OBJETIVOS	16
Objetivo Geral.....	16
Objetivos Específicos	16
FUNDAMENTAÇÃO TEÓRICA	16
Aprendizado de máquina	16
Métricas de avaliação contínua	19
Métricas de avaliação categórica.....	20
Sobreajuste	23
Validação cruzada	23
Árvores de decisão e Florestas aleatórias	25
Descritores Moleculares e Fingerprints	27
QSAR.....	28
REFERÊNCIAS	30
 CAPÍTULO II.....	 31
AmIActive (AIA): A Large-scale QSAR Based Target Fishing and Polypharmacology Predictive Web Tool.....	32
 CONCLUSÕES.....	 46
ANEXOS I Produção Científica.....	47
ANEXOS II Eventos e Congressos.....	49

CAPÍTULO I

INTRODUÇÃO

O planejamento de fármacos auxiliado por computador (*Computer-Aided Drug Design*, CADD) consiste no conjunto de metodologias computacionais empregadas para apoiar, acelerar e racionalizar o processo de descoberta e desenvolvimento de novos compostos bioativos. Na química medicinal moderna, o CADD desempenha um papel central ao permitir a análise sistemática de grandes volumes de dados químicos e biológicos, reduzindo custos experimentais, tempo de desenvolvimento e taxas de falha associadas às etapas iniciais da descoberta de fármacos [1]. As abordagens de CADD podem ser amplamente classificadas em métodos baseados em estrutura, que exploram informações tridimensionais de alvos biológicos, e métodos baseados em ligantes, que utilizam dados experimentais de atividade biológica associados a estruturas químicas conhecidas [2].

Dentro desse contexto, os métodos baseados em ligantes ganharam especial relevância com o crescimento exponencial de bases de dados públicas de atividade biológica, como o ChEMBL [3], que disponibilizam milhões de medições experimentais obtidas a partir de ensaios farmacológicos. Essas abordagens fundamentam-se na premissa de que moléculas estruturalmente semelhantes tendem a apresentar perfis biológicos semelhantes, permitindo a construção de modelos capazes de inferir relações quantitativas entre estrutura química e atividade biológica. Entre essas metodologias, destacam-se os modelos QSAR (*Quantitative Structure–Activity Relationship*), amplamente utilizados para prever propriedades farmacológicas e orientar decisões em química medicinal [4].

Dessa forma, ao integrar dados experimentais, descritores moleculares, algoritmos de aprendizado de máquina e estratégias rigorosas de validação, as abordagens CADD constituem um conjunto de ferramentas indispensáveis na química medicinal contemporânea [5]. Em especial, os modelos QSAR representam um elo fundamental entre a estrutura química e a resposta biológica, viabilizando a predição racional de atividades farmacológicas e contribuindo de maneira decisiva para o avanço de estratégias modernas de descoberta de fármacos. Associado a isto, graças ao desenvolvimento e adoção exponencial da internet nos últimos anos múltiplas ferramentas web foram construídas com o objetivo de universalizar a aplicação e utilização desses algoritmos e ferramentas.

Dado este contexto, este trabalho tem como objetivo o desenvolvimento de uma ferramenta CADD publicamente disponível na internet intitulada de AmIActive (AIA). Neste capítulo, os principais conceitos e definições fundamentais para a compreensão do funcionamento da ferramenta desenvolvida são introduzidos. No capítulo seguinte, o desenvolvimento do sistema AIA é detalhado.

OBJETIVOS

Objetivo Geral

Desenvolver um sistema robusto, publicamente disponível e de fácil uso, com a capacidade de prever múltiplas atividades biológicas distintas para uma ou mais moléculas.

Objetivos Específicos

- Selecionar, filtrar e limpar dados biológicos oriundos da base de dados ChEMBL;
- Treinar modelos *Random Forest* nos dados obtidos e selecionar modelos estatisticamente significativos;
- Realizar ajustes nos hiper-parâmetros dos modelos obtidos;
- Validar os modelos obtidos através do procedimento de validação cruzada;
- Calcular métricas de desempenho dos modelos obtidos (acurácia, precisão, sensibilidade, etc.);
- Desenvolver interface web amigável para exposição dos modelos aos usuários finais;
- Lançar a plataforma no endereço web amiactive.ccen.ufpb.br.

FUNDAMENTAÇÃO TEÓRICA

Aprendizado de máquina

O aprendizado de máquina, também conhecido como aprendizado estatístico, é uma subárea de estudo da estatística cujo o objetivo é aproximar uma função desconhecida através de um conjunto de variáveis de entrada, também chamadas de variáveis preditoras ou variáveis independentes. Mais formalmente, tendo-se uma variável de interesse, também chamada de variável dependente, sendo esta qualquer característica mensurável que o pesquisador tenha interesse em prever, e tendo-se variáveis que o pesquisador acredita estarem de alguma maneira relacionadas à variável de interesse, os algoritmos de aprendizado de máquina buscam aproximar a função

desconhecida que relaciona as variáveis que o pesquisador acredita serem relevantes (variáveis independentes) até a variável de interesse (variável dependente) em questão [6].

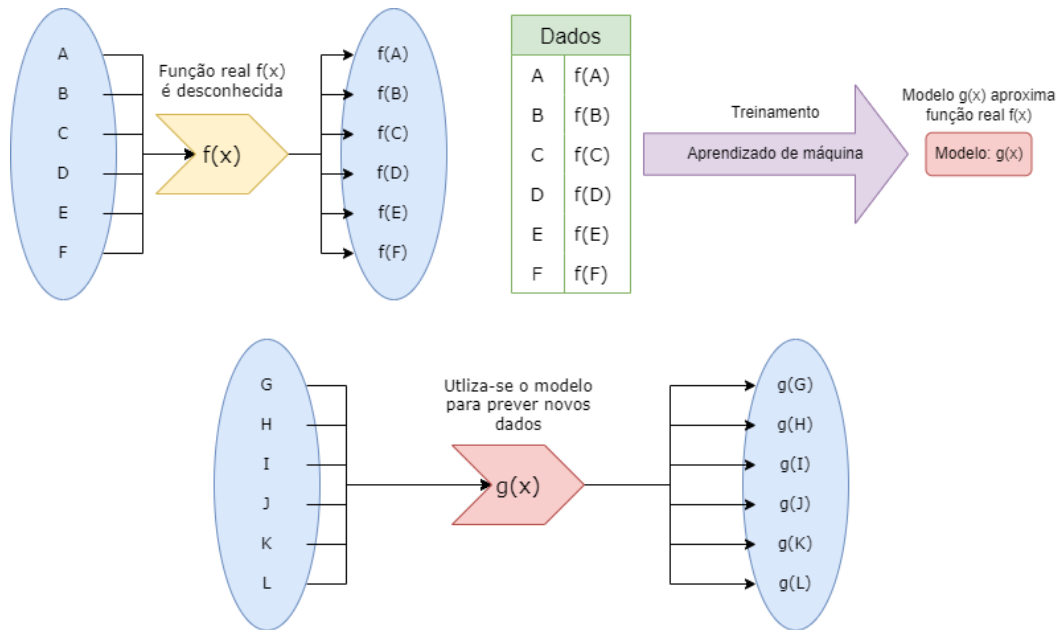


Figura 1: Esquematização do funcionamento do aprendizado de máquina.

Na terminologia do aprendizado de máquina, a instância específica da aplicação de um algoritmo para a aproximação de uma função de interesse é chamada de modelo. Para construir-se um modelo, é necessário ter dados coletados a respeito da característica de interesse que se deseja prever. Mais precisamente, é preciso ter coletado instâncias individuais da variável de interesse e de suas variáveis dependentes, de maneira a ter múltiplos exemplos da relação entre as variáveis independentes e a variável de interesse em questão. Feito isto, aplicando-se os algoritmos de aprendizado de máquina aos dados em questão, uma função que mapeia de maneira aproximada as variáveis independentes até a variável de interesse será construída. O processo de aplicação de um algoritmo de aprendizado de máquina a um conjunto de dados específico é denominado de treinamento (Figura 1). O modelo treinado contém toda a lógica e as regras para mapear variáveis de entrada com valores específicos até suas saídas correspondentes. Desta maneira, em posse de um modelo de aprendizado de máquina, um pesquisador pode prever instâncias individuais de casos de interesse utilizando as variáveis independentes associadas ao caso específico.

Variáveis de interesse em tarefas de aprendizado de máquina podem ser divididas em variáveis de interesse categóricas, quando a variável de interesse em questão só pode assumir valores discretos dentro de uma faixa estrita, e variáveis de interesse contínuas, quando a variável de interesse pode assumir qualquer valor numérico dentro de uma faixa contínua. Quando a tarefa possui uma variável de interesse contínua o problema de aprendizado de máquina é denominado de problema de regressão e o modelo treinado é chamado de regressor. De maneira análoga, quando a tarefa possui uma variável de interesse categórica o problema de aprendizado de máquina é denominado de problema de classificação e o modelo treinado é chamado de classificador e os valores discretos que a variável de interesse pode assumir são chamados de classes. Em especial, quando a variável de interesse de um problema de classificação possui apenas dois valores possíveis, o problema é chamado de classificação binária (Figura 2).

Embora extremamente úteis, os modelos de aprendizado de máquina são fundamentalmente aproximações das funções reais desconhecidas, treinados a partir de dados experimentais que já possuem por si só erros aleatórios, desta maneira, sempre haverá algum grau de erro e, consequentemente, incerteza envolvida nas previsões retornadas pelo modelo. Considerando isto, é de vital importância que o modelo seja prontamente avaliado em relação às suas capacidades preditivas de maneira a entender até que ponto pode se confiar nas previsões geradas por ele. Para se avaliar o desempenho dos modelos treinados, utiliza-se de métricas de desempenho distintas, que variam de acordo com o tipo da variável de interesse (categórica/contínua). Desta maneira, fazendo uso de dados em que a variável de interesse foi experimentalmente mensurada, utiliza-se o modelo treinado para prever a variável de interesse nestes dados experimentalmente mensurados. Através da comparação entre as previsões do modelo e os dados da variável de interesse experimentalmente determinados, é possível estimar as capacidades preditivas do modelo quando este for utilizado em dados em que a variável de interesse não foi experimentalmente mensurada, sendo, portanto, desconhecida. Uma discussão mais detalhada das métricas de desempenho segue mais adiante.

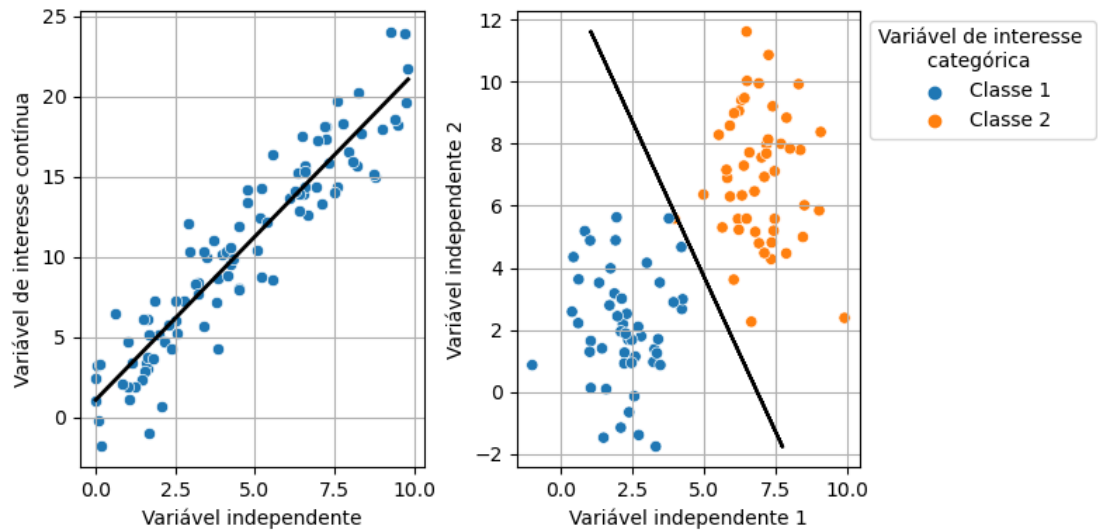


Figura 2: Exemplos de problemas de regressão e classificação.

Apesar de certas terminologias e algoritmos serem de desenvolvimento e disseminação recente, o aprendizado de máquina tem origens históricas que remontam aos primórdios da estatística como ciência, tendo como origem, os modelos lineares desenvolvidos no ramo da estatística denominado de análise multivariada de dados, tais como: Regressão linear, análise discriminante linear, regressão logística, dentre outros.

Métricas de avaliação contínua

MAE

O desvio médio absoluto (*mean absolute error; MAE*) (Equação 1), é uma métrica de avaliação de regressão. Esta métrica consiste na média simples dos erros absolutos do modelo. Mais especificamente, para todos os dados, calcula-se o valor absoluto da diferença entre o valor real da característica de interesse e a predição dada pelo modelo e posteriormente, tira-se a média de todos esses valores. O desvio médio absoluto fornece o erro médio do modelo dentro da escala da variável de interesse [7].

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}. \quad (1)$$

MSE

O desvio médio quadrado (*mean squared error; MSE*) (Equação 2), é uma métrica de avaliação de regressão muito similar ao desvio absoluto médio. O desvio médio quadrado é calculado da mesma maneira que o desvio absoluto médio, no entanto, ao invés de calcular-se o valor absoluto das diferenças, calcula-se o quadrado das diferenças. Ao calcular o quadrado das diferenças, esta métrica passa a não estar mais na escala original da variável de interesse, porém, ao tomar o quadrado das diferenças, essa métrica tende a ser mais sensível a erros extremos, sendo mais conservativa no que tange a estimação do erro do modelo [7].

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (2)$$

RMSE

A raiz do desvio quadrado médio (*root mean squared error; RMSE*) (Equação 3), é uma métrica de avaliação de regressão derivada do desvio médio quadrado. Ao calcular-se a raiz quadrada do desvio médio quadrado, converte-se o desvio dos dados a escala original, permitindo assim, uma interpretação direta dos desvios sumarizados com esta métrica, ao mesmo tempo em que, assim como no desvio médio quadrado, penaliza-se mais desvios extremos [7].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - p_i)^2} \quad (3)$$

Métricas de avaliação categórica

Acurácia

A Acurácia (Equação 4), é uma métrica de avaliação de classificação caracterizada por estimar de maneira global o desempenho do modelo. A Acurácia é a proporção entre a quantidade de previsões, negativas ou positivas, que foram corretamente classificadas

pelo modelo e quantidade dados que foram utilizados para a avaliação do modelo. Apesar de ser uma métrica que avalia simultaneamente o desempenho do modelo em instâncias positivas e negativas, esta métrica é afetada consideravelmente em casos em que as classes não estão balanceadas [7].

$$\textbf{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (4)$$

Precisão

A Precisão (Equação 5), é uma métrica de avaliação de classificação caracterizada por estimar a probabilidade de que uma determinada predição do modelo esteja correta. Esta métrica representa a capacidade de um modelo estar correto em uma predição individual e pode ser calculada para cada uma das classes presentes, não sendo desta maneira, uma métrica de desempenho global [7].

$$\textbf{Precisão} = \frac{VP}{VP + FP} \quad (5)$$

Sensibilidade

A Sensibilidade (Equação 6), é uma métrica de avaliação de classificação caracterizada por estimar a probabilidade de que uma instância individual será corretamente classificada pelo modelo. Essa métrica representa a capacidade de um modelo identificar uma instância de uma classe específica, sendo desta maneira, a métrica complementar a precisão. Assim como a precisão, a Sensibilidade pode ser calculada para cada uma das classes, no entanto, em um contexto de classificação binária, a Sensibilidade se refere a classe positiva, de maneira que a mesma métrica calculada para a classe negativa recebe outro nome (ver Especificidade a seguir) [7].

$$\textbf{Sensibilidade} = \frac{VP}{VP + FN} \quad (6)$$

Especificidade

A Especificidade (Equação 7), é uma métrica de avaliação de classificação caracterizada por ser a Sensibilidade da classe negativa. Deste modo, esta métrica representa a capacidade de um modelo identificar corretamente uma instância da classe negativa [7].

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (7)$$

Score F

O Score F (Equação 8), é uma métrica de avaliação de classificação caracterizada por ser a média harmónica entre a precisão e a sensibilidade, representando desta maneira ambas as métricas simultaneamente. Dado isto, esta métrica representa tanto a capacidade de um modelo fazer previsões corretas, como também a capacidade de identificar corretamente instâncias de uma classe específica. Apesar de sintetizar a precisão e a sensibilidade em uma métrica única, como precisão e sensibilidade são calculadas para classes diferentes, o Score F também é calculado para classes individuais, tornando-o uma métrica não global [7].

$$\text{Score F} = 2 \cdot \frac{\text{precisão} \times \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}} \quad (8)$$

Coeficiente de correlação de Matthews (CCM)

O Coeficiente de Correlação de Matthews (Equação 9), é uma métrica de avaliação de classificação caracterizada por ser a correlação entre as previsões realizadas pelo modelo e as classes reais de cada instância dos dados presentes. O CCM equivale a correlação de Pearson quando as variáveis sendo observadas são binárias. Deste modo, o CCM representa o desempenho do modelo em ambas as classes simultaneamente sendo desta maneira, uma métrica de desempenho global. Além disto, o CCM não é afetado por classes desbalanceadas. Devido a isto, o CCM é considerado uma das melhores métricas numéricas para avaliação de classificadores [8].

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}} \quad (9)$$

Sobreajuste

Em dados reais, sempre haverá uma variação aleatória entre a relação das variáveis independentes e a variável de interesse. Isto pode ocorrer pelo fato de que outras variáveis independentes não utilizadas no modelo serem necessárias para modelar toda a variância da variável de interesse ou por que a própria variável de interesse possui um caráter inerentemente estocástico. O sobreajuste ou *overfitting* é um fenômeno que ocorre quando um modelo de aprendizado de máquina modela não apenas a relação real entre as variáveis independentes e a variável de interesse, mas também as variações aleatórias presentes nos dados (Figura 3). Em consequência disto, o modelo passa a fazer previsões considerando estas variações aleatórias como reais quando elas de fato não são, danificando consideravelmente as capacidades preditivas do modelo em questão. Para reduzir o sobreajuste de modelos individuais, utiliza-se procedimentos de regularização, redução de variância e diminuição da complexidade do modelo [9].

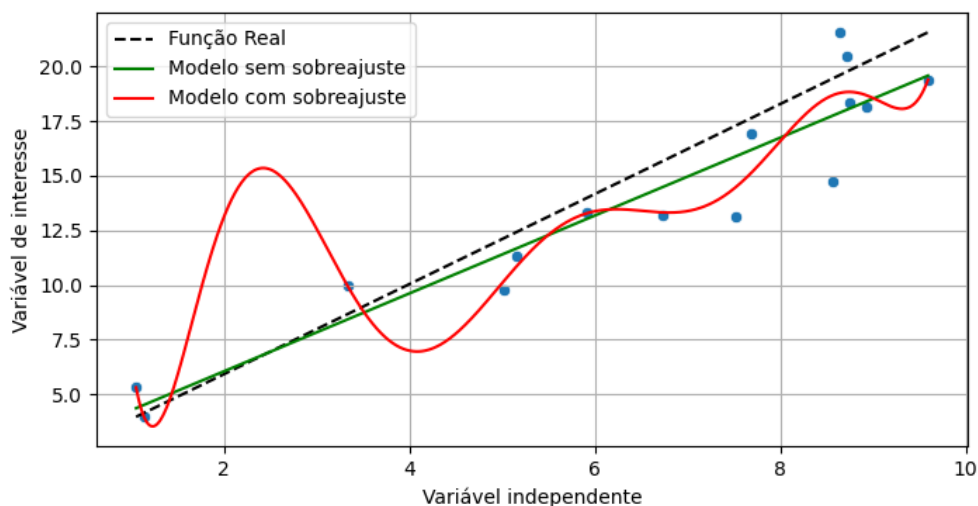


Figura 3: Exemplo de sobreajuste.

Validação cruzada

Apesar de algoritmos de aprendizado de máquina modernos possuírem uma série de mecanismos para reduzir o sobreajuste, ainda é esperado algum grau de modelação das flutuações aleatórias presentes nos dados e consequentemente, ainda é esperado algum grau de sobreajuste. Além de piorar a capacidade preditiva, o sobreajuste gera problemas na estimação da performance do modelo. Mais especificamente, devido a modelagem das flutuações aleatórias nos dados de treino, a avaliação da performance do modelo nos mesmos dados em que ele foi treinado irá levar invariavelmente a uma estimativa muito

otimista do desempenho real do modelo. Isto acontece por que já que as flutuações aleatórias que estão presentes nos dados de treino são as mesmas que o modelo capta através do sobreajuste, o modelo irá “predizer” corretamente essas flutuações aleatórias apesar de elas não serem de fato reais. Em consequência disto, o modelo terá o desempenho aparentemente melhor nos dados em que ele foi treinado do que em dados que não foram utilizados durante o treinamento.

Considerando isto, para obter-se uma estimativa não enviesada da capacidade preditiva de um modelo, é necessário avaliar-se a sua performance em dados que não foram utilizados para realizar o treinamento do modelo. Quando não se possui dados especificamente coletados para avaliação do modelo, uma das possíveis abordagens é a divisão dos dados em um subconjunto de treinamento e um subconjunto de teste. Desta maneira, obtém-se uma estimativa não enviesada da performance do modelo treinando-o com o subconjunto de treino e avaliando-o com o subconjunto de teste. Embora simples, a divisão entre as instâncias de dados individuais para cada subconjunto gera ambiguidade e variabilidade na estimativa de performance do modelo, tendo em vista que para conjuntos de dados de tamanho moderado existem uma infinidade de maneiras de se organizar os dados em subconjuntos diferentes. Em função disto, divisões diferentes de subconjuntos de treino e teste podem ter estimativas de performance diferentes, já que tanto os dados utilizados para treinar o modelo quanto os dados utilizados para testar a performance, são diferentes.

Em função do exposto, para reduzir esta variabilidade de divisão, utiliza-se o procedimento de validação cruzada (Figura 4). Este procedimento consiste em dividir o conjunto de dados em K partições (usualmente 5 ou 10), com cada partição representando um subconjunto dos dados originais. Para cada partição, um modelo será treinado com as partições remanescentes e avaliado na partição selecionada. Ao final deste procedimento, tira-se a média do desempenho considerando os resultados obtidos em cada partição. A validação cruzada, diminui consideravelmente a variabilidade associada a divisão arbitrária do conjunto de dados devido ao fato de que a cada etapa do processo, o modelo é treinado com um subconjunto de treino e um subconjunto de teste diferente. Em função disto, a validação cruzada é método de avaliação padrão para modelos de aprendizado de máquina, tanto na academia quanto na indústria [10].

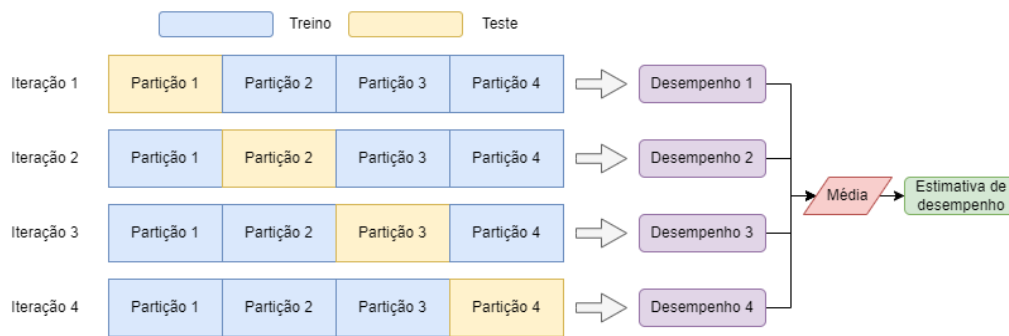


Figura 4: Esquematização do procedimento de validação cruzada.

Árvores de decisão e Florestas aleatórias

As árvores de decisão, são um tipo de algoritmo de aprendizado de máquina hierárquico que podem gerar tanto regressores quanto classificadores [11]. Esse algoritmo baseia-se na divisão sucessiva do conjunto de dados de treino em subconjuntos cada vez menores até que os subconjuntos atinjam um determinado grau de consistência em relação a variável de interesse e portanto, sejam consideravelmente similares. Desta maneira, ao se mapear uma instância de interesse até uma dessas subdivisões, pode-se obter uma predição da variável de interesse (Figura 5).

Árvore de decisão

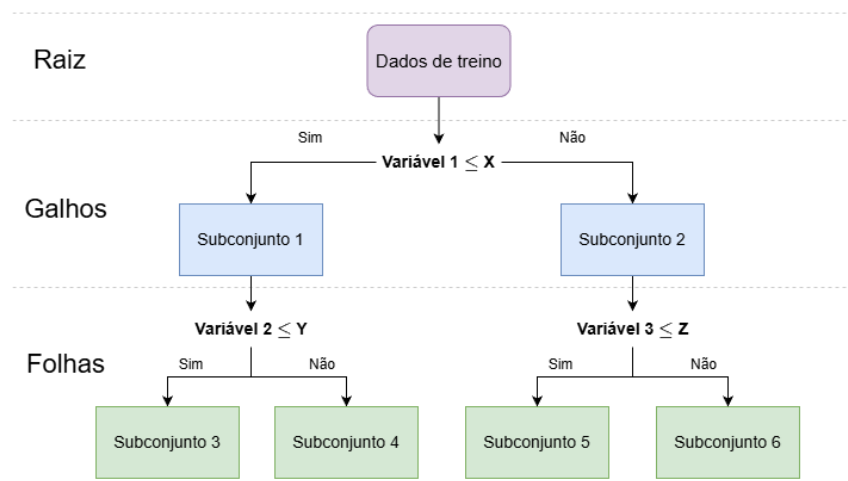


Figura 5: Esquematização de uma árvore de decisão

As árvores de decisão são treinadas a partir de um algoritmo iterativo chamado de *Recursive Binary Splitting*. Inicialmente, todos os dados de treino estão no mesmo subconjunto de dados, denominado de raiz (*root*) da árvore. A cada iteração, o algoritmo testa todas as divisões possíveis dos dados considerando todas as variáveis independentes e seleciona a variável independente e a divisão que maximiza um determinado critério de

eficácia (Desvio médio quadrado, índice de gini, entropia, etc.). Uma vez feita essa subdivisão o processo é repetido em cada uma das subdivisões resultantes até que determinado critério de parada seja atingido. As subdivisões intermediárias são chamadas de galhos (*branches*) e as subdivisões finais utilizadas para a predição do modelo são chamadas de folhas (*leaves*). As árvores de decisão são conhecidas por serem modelos não paramétricos altamente não lineares e com capacidade de aprender padrões complexos. No entanto, essa grande flexibilidade os torna particularmente susceptíveis a problemas de sobreajuste severos de maneira que sua aplicação acaba se tornando muito limitada [12].

Para atenuar os problemas de sobreajuste das árvores de decisão Leo Breiman criou o algoritmo Florestas aleatórias (*Random Forest*) em 2001 [13]. Esse algoritmo baseia-se na premissa de que uma predição baseada na votação de um grupo de múltiplos modelos (chamado de *ensemble*) pode ser muito superior a qualquer predição individual de um determinado modelo do grupo, desde que os modelos base do grupo sejam consideravelmente diferentes uns dos outros. As florestas aleatórias utilizam árvores de decisão como modelos base para a formação do grupo (Figura 6).

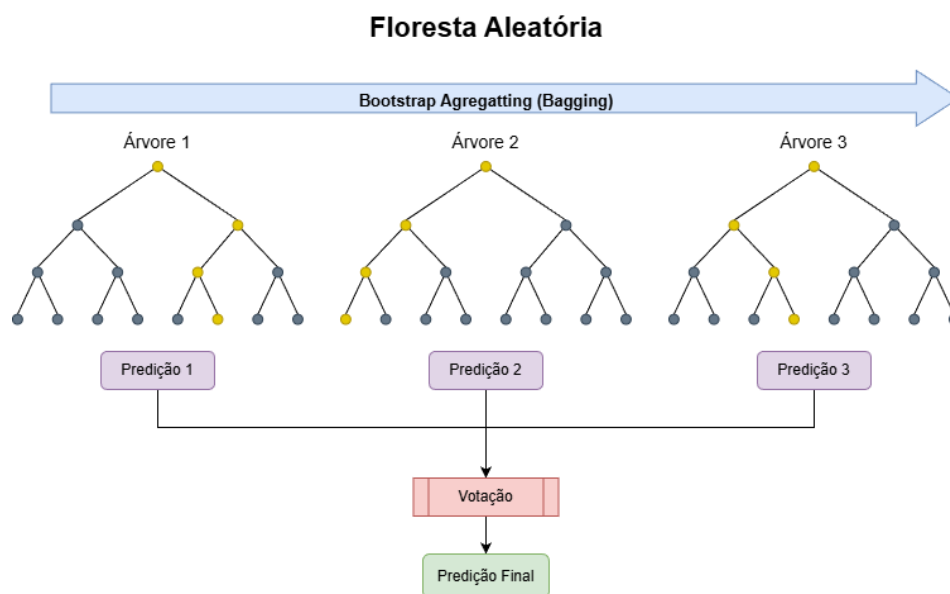


Figura 6: Esquematização de uma floresta aleatória

Os modelos floresta aleatória são treinados utilizando um processo chamado de *Bootstrap Aggregating* também chamado de *Bagging*. Neste algoritmo um número de predeterminado de árvores de decisão é treinado com os dados de treino seguindo o mesmo algoritmo delineado anteriormente, mas com duas modificações. Em primeiro

lugar, cada árvore é treinada com uma amostra com reposição dos dados de treino originais. Desta maneira, cada árvore individual é treinada com um conjunto de dados parcialmente diferente. Em segundo lugar, durante o processo de “crescimento” da árvore, ao invés de considerar todas as variáveis independentes para a realização da melhor divisão, a cada etapa de divisão, somente um subconjunto aleatório de variáveis independentes é considerado. Esse procedimento aumenta consideravelmente a diversidade das árvores de decisão treinadas, aumentando assim a independência entre suas previsões. A soma desses procedimentos possui o efeito de diminuir muito consideravelmente a variância dos modelos individuais e por consequência, o sobreajuste resultante. Deste modo, as florestas aleatórias possuem as mesmas vantagens das árvores de decisão com o benefício adicional de serem muito menos afetadas pelo sobreajuste [14].

Descritores Moleculares e *Fingerprints*

Descritores moleculares são representações numéricas de determinadas características das moléculas [15]. De maneira geral, os descritores podem ser classificados de acordo com o tipo de informação extraída das moléculas (Físico-químicos, Topológicos ou Geométricos), bem como pelo grau de dimensionalidade espacial dessa informação (1D, 2D, 3D, etc.). O objetivo dos descritores é codificar características moleculares relevantes. Deste modo, a escolha de um determinado tipo de descritor está intimamente relacionada aos fatores moleculares importantes para o problema em questão [16].

Os *Fingerprints* moleculares são um tipo de descritor molecular topológico 2D baseados na conectividade dos átomos dentro de uma molécula [17]. Essa classe de descritores mapeia subestruturas moleculares em uma representação vetorial composta por *bits* (1 ou 0) de maneira que um *bit* com o valor de 1 representa a presença de uma determinada subestrutura enquanto um *bit* com o valor de 0 representa a sua ausência da molécula em questão (Figura 7). O mapeamento das subestruturas pode ser baseado em um dicionário de subestruturas pré-definido, como no caso dos *Fingerprints* MACCS (*Molecular ACCess System*) [18], ou baseado em um algoritmo dinâmico como no caso dos *Fingerprints* ECFP (*Extended-Connectivity Fingerprints*) [19].

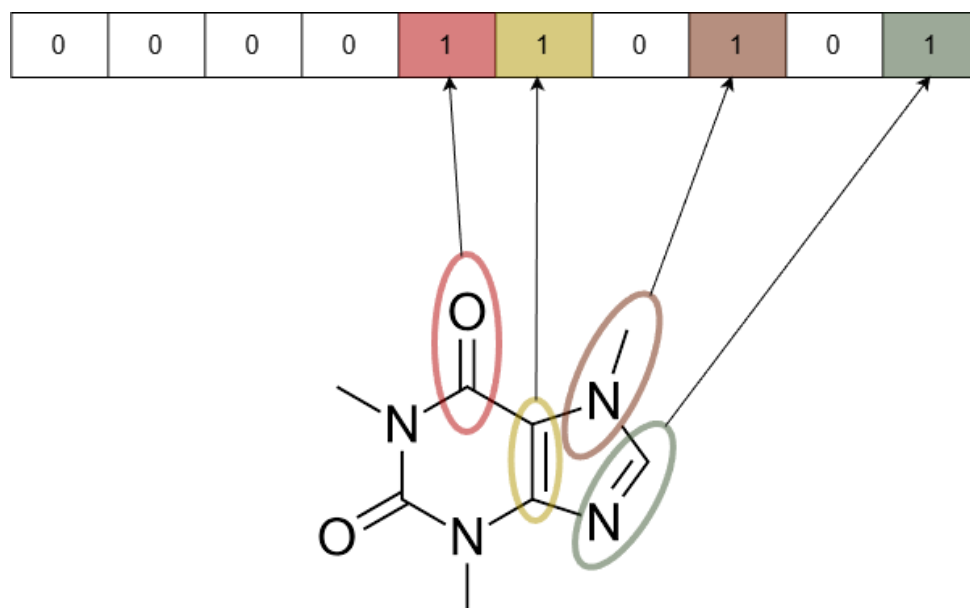


Figura 7: Esquemática de um *Fingerprint* genérico

QSAR

Os modelos QSAR, acrônimo para *Quantitative Structure-Activity Relationship*, são modelos de aprendizado de máquina treinados para prever resultados de ensaios farmacológicos de potência baseado na estrutura química das moléculas presentes no conjunto de dados. Mais precisamente, através de descritores moleculares, extrai-se informações estruturais das moléculas que compõe o conjunto de dados experimental. Uma vez calculados os descritores moleculares, constrói-se um modelo de aprendizado de máquina utilizando os descritores como variáveis independentes e uma métrica de desempenho em um ensaio biológico específico como variável de interesse. Deste modo, o modelo treinado é capaz de prever a atividade biológica de uma molécula baseado em sua estrutura molecular (Figura 8).

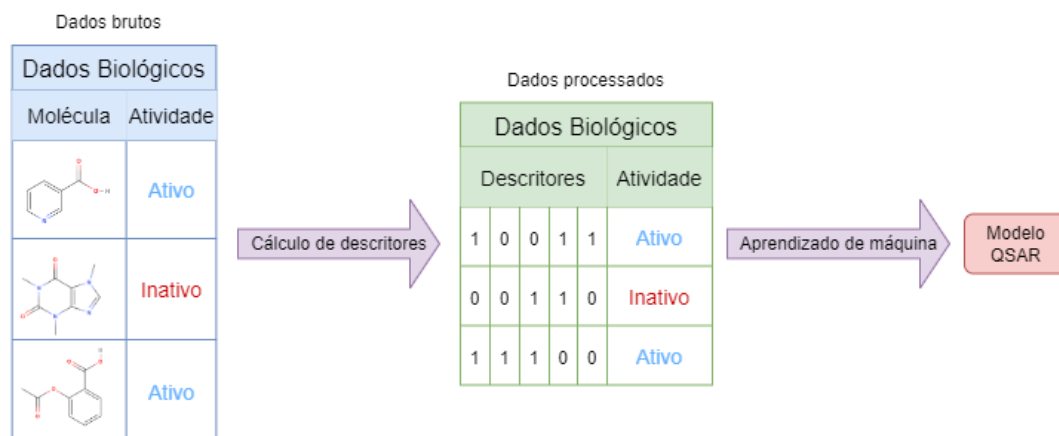


Figura 8: Esquematização do processo de treinamento de modelos QSAR.

Os modelos QSAR são baseados no princípio SAR (*Structure-Activity Relationship*). Este princípio estabelece que moléculas similares tendem a ter atividades similares em um mesmo tipo de atividade biológica. A racionalidade deste princípio baseia-se no fato de que a atividade dos fármacos está intimamente relacionada a as interações moleculares entre os fármacos e os seus alvos moleculares de interesse. Considerando que a quantidade e a natureza das interações moleculares de uma molécula pequena e o seu alvo de interesse são governadas pela estrutura atômica da mesma, é possível postular que, de maneira geral, moléculas similares em termos de composição atômica e conectividade entre os átomos terão atividades correlacionadas [20].

REFERÊNCIAS

- [1] OOMS, Frédéric. Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. *Current medicinal chemistry*, v. 7, n. 2, p. 141-158, 2000.
- [2] SHARMA, Vidushi; WAKODE, Sharad; KUMAR, Hirdesh. Structure-and ligand-based drug design: Concepts, approaches, and challenges. *Chemoinformatics and bioinformatics in the pharmaceutical sciences*, p. 27-53, 2021.
- [3] ZDRAZIL, Barbara et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, v. 52, n. D1, p. D1180-D1192, 2024.
- [4] HANSCH, Corwin. Quantitative structure-activity relationships and the unnamed science. *Accounts of chemical research*, v. 26, n. 4, p. 147-153, 1993.
- [5] CHETRY, Anup Basnet; OHTO, Keisuke. From molecules to data: the emerging impact of chemoinformatics in chemistry. *Journal of Cheminformatics*, v. 17, n. 1, p. 121, 2025.
- [6] VAPNIK, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10, n. 5, p. 988-999, 1999.
- [7] HANDELMAN, G. S. et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212, n. 1, p. 38-43, 2019.
- [8] CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21, n. 1, p. 1-13, 2020.
- [9] YING, X., 2019, An overview of overfitting and its solutions. IOP Publishing. 022022.
- [10] STONE, M. Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics*, 9, n. 1, p. 127-139, 1978.
- [11] BREIMAN, Leo et al. *Classification and regression trees*. Chapman and Hall/CRC, 2017.
- [12] SCHAFFER, Cullen. When does overfitting decrease prediction accuracy in induced decision trees and rule sets?. In: *European Working Session on Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991. p. 192-205.
- [13] BREIMAN, Leo. Random forests. *Machine learning*, v. 45, n. 1, p. 5-32, 2001.
- [14] HALABAKU, Erblin; BYTYÇI, Eliot. Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests. *Intelligent Automation & Soft Computing*, v. 39, n. 6, 2024.
- [15] NIAZI, Sarfaraz K.; MARIAM, Zamara. Recent advances in machine-learning-based chemoinformatics: a comprehensive review. *International Journal of Molecular Sciences*, v. 24, n. 14, p. 11488, 2023.
- [16] KHAN, Asad U. et al. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug discovery today*, v. 21, n. 8, p. 1291-1302, 2016.
- [17] CERETO-MASSAGUÉ, Adrià et al. Molecular fingerprint similarity search in virtual screening. *Methods*, v. 71, p. 58-63, 2015.
- [18] DURANT, Joseph L. et al. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences*, v. 42, n. 6, p. 1273-1280, 2002.
- [19] ROGERS, David; HAHN, Mathew. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, v. 50, n. 5, p. 742-754, 2010.
- [20] MURATOV, E. N. et al. QSAR without borders. *Chemical Society Reviews*, 49, n. 11, p. 3525-3564, 2020.

CAPÍTULO II

AmIActive (AIA): A Large-scale QSAR Based Target Fishing and Polypharmacology Predictive Web Tool

Luis Felipe de Morais Melo^{1,2}, Uilames de Assis Ferreira¹, Renan Paiva Oliveira Costa¹, Luciano Pereira de Oliveira Filho¹, Eduardo Henrique Pessoa Alves¹, Luciana Scotti², Marcus Tullius Scotti^{1,2}

1- Cheminformatics Laboratory, Department of Chemistry, Center for Exact and Natural Sciences (CCEN), Federal University of Paraiba (UFPB), Campus I Lot. Cidade Universitaria, Castelo Branco, João Pessoa, Paraiba, Brazil

2- Postgraduate Program in Natural and Synthetic Bioactive Products (PgPNSB), Federal University of Paraiba (UFPB), Campus I Lot. Cidade Universitaria, Castelo Branco, João Pessoa, Paraiba, Brazil

Artigo publicado na revista *Journal of Molecular Biology (JMB)*, 2025, 437, 15.

DOI: <https://doi.org/10.1016/j.jmb.2025.169090>

Fator de Impacto: 4.5

Abstract: Here, we introduce AmIActive (AIA), a QSAR-based web tool for biological activity prediction and target fishing. The AIA system uses QSAR models trained with biological activity data from the ChEMBL database and currently covers single proteins, protein complexes, protein families, cell lines, organisms, and tissues. The system contains 3,239 models corresponding to 2,277 distinct targets of interest. These models are Random Forest classifiers trained with circular fingerprint descriptors, and all models have clearly defined applicability domains. The system is easy to use, with a friendly and intuitive interface. The results can be downloaded in CSV format and include predictive information, details about the target, validation metrics for the models, and the active/inactive threshold used to train the classifier. The AIA system is available free of charge and can be accessed at the following link: <https://amiactive.ccen.ufpb.br/>.

Highlights:

- AmIActive (AIA) is a QSAR target fishing web application with diverse targets;
- The AIA system has 3239 QSAR models trained on datasets from the ChEMBL database;
- The models system are Random Forest classifiers trained with ECFP4 fingerprints;
- With AIA, a researcher can prioritize more promising pharmacological activities;
- AIA is easy to use and can be accessed freely at: <https://amiactive.ccen.ufpb.br/>

Introduction

Computer Aided Drug Design (CADD) is a research field that studies the application of computer systems and algorithms in the development of hits, leads and drugs for specific targets of interest [1]. CADD techniques can be categorized into ligand-based techniques, when only information about the ligand of interest is used in the prediction process (similarity search, pharmacophore modeling, etc.) or structure-based techniques, when the structure of the target is used in the prediction process (molecular docking, molecular dynamics, among others.) [2].

CADD software and tools have been widely adopted with a recent breakthrough in protein structure prediction with AlphaFold [3]. AlphaFold is an advanced artificial intelligence system developed by DeepMind that predicts the three-dimensional structure of proteins from their amino acid sequences with considerable accuracy. The impact of this system has been tremendous, sparking the development of several tools, web servers and databases based on it [4-6]. In particular, AlphaFold has significantly expanded the use of molecular docking. Previously, if a high-quality experimental 3D structure of the target was unavailable, users could only rely on homology modeling [7].

Quantitative Structure Activity Relationship (QSAR) is a ligand-based, CADD technique that works by learning correlation patterns between the structure of molecules and their biological activities. This technique has achieved a high adoption in the chemoinformatics, computational chemistry and medicinal chemistry communities [8]. One of the disadvantages of QSAR modeling in comparison to other CADD techniques such as molecular docking is the need of already existing biological activity datasets in order to train the desired QSAR models [9].

Although still relevant, this disadvantage has been significantly reduced by the increasing number of publicly available biological activity data in the literature. This is evidenced by the increasing number of biological activities present in public databases such as ChEMBL [10], BindingDB [11], among others. As such, coupled to this increase in data and advancements in computing power and machine learning algorithms, multiple public available QSAR based web tools have been developed to further facilitate the use of these methods for non-specialists [12].

The first of such tools was the PASS (Prediction of Activity Spectra for Substances) web system, developed by Poroikov and colleagues and in use ever since

1999 [13]. This system is composed of more than 5000 biological activities and uses models based on a modified version of the naïve bayes classifier trained with multilevel neighborhood of atoms (MNA) descriptors as features. More recently, the group has also published a specific version of the aforementioned system called PASS-targets [14], which is trained in a filtered version of the ChEMBL database. Another tool developed by the same group is GUSAR (General Unrestricted Structure-Activity Relationships) which is similar in scope but predicts quantitative measures of biological activity (IC50, EC50, etc.). GUSAR currently has 3663 biological activities, and it uses a consensus of self-consistent regression (SCR), nearest neighbor evaluation and radial basis function artificial neural networks (RBF ANN) models trained with a combination of "Quantitative Neighborhoods of Atoms" (QNA) descriptors, physiochemical descriptors and the probabilistic output of the previously mentioned PASS system.

Another public available QSAR based web tool is PLATO (Polypharmacology pLATform predictiOn) developed by Ciriaco and colleagues [15]. PLATO is a quantitative predictive tool built with models trained with the biological activity data from the ChEMBL database. The models of this system use two custom predictive algorithms made specifically for target fishing using a set of 13 descriptors as features. As of the writing of this article, this system is made of 6004 distinct biological activities.

SwissTargetPrediction is another web based predictive tool that was developed by Daina and colleagues [16]. The models of this system were trained with biological activity data from the ChEMBL database and it currently holds 3068 models of distinct biological activities. These models are logistic regression classifiers trained to weigh a 3D similarity measure (Manhattan distance) calculated with ES5D descriptors and a 2D similarity measure (Tanimoto similarity) calculated with FP2 descriptors between a query molecule and the molecules of known biological activity in order to make a combined similarity score.

The previous paragraphs are by no means an extensive list of the tools in this ecosystem with plenty of other popular tools also being available [17-18]. Although drug potency is one of the most important factors to be optimized in a drug discovery pipeline, other factors such as pharmacokinetics and toxicology also play a critical role. Therefore, a complete drug discovery pipeline needs to consider these additional factors even at the *in silico* level [19-21]. One of the popular tools in this ecosystem is SwissADME [22], a web tool developed by the same group that built

SwissTargetPrediction. This tool can predict physiochemical, pharmacokinetic, toxicology and drug likeness properties for individual molecules.

In this work we have developed another QSAR based web tool called AmIActive (AIA) which we hope will add further value to the already existing ecosystem of such tools. The AIA system is built with the biological activity data from the ChEMBL database and its models are Random Forest classifiers trained with circular fingerprint descriptors. the AIA system models cover molecular and non-molecular targets such as single proteins, protein complexes, protein families, cell-lines, organisms and tissues. In addition to a broad set of validation metrics (accuracy, precision, recall, etc.), we also provide the active/inactive activity threshold used to train the underlying model, helping researchers decide whether pursuing *in vitro* testing is worthwhile based on their expectations. In addition to the predicted activity, we also provide the applicability domain calculation for each prediction. This ensures that our system complies with (Q)SAR Model Principle 3 (OECD, 2007), which states that “...a valid model is associated with a defined applicability domain (AD).”. Finally, to the best of our knowledge, the AIA system is the only tool of its kind in Brazil, making it an important milestone in Brazilian science. The AIA system is available at the link: <https://amiactive.ccen.ufpb.br/>.

Materials and methods

ChEMBL data selection

The datasets used for QSAR model building in this work were obtained from the ChEMBL30 database. Succinctly, a downloaded raw sql dump of the ChEMBL database was read in a local PostgreSQL database. In sequence, we selected all targets that were either single proteins, cell-lines, protein families, protein complexes, organisms or tissues. Finally, we separated the activities of each target into different datasets on the basis of the standard type (IC₅₀, K_i, etc.), where each generated dataset represents a target-activity type pair.

Dataset cleaning and standardization

Datasets with less than 30 molecules, with units that could not be converted to some form of molar representation (nM, uM, etc) and with standard types that were not based on some form of dose-response curve (IC₅₀, EC₅₀, etc.) were removed from the

work. The remaining datasets were subsequently cleaned using the best practices established in the literature [23]. Briefly, for each dataset, mixtures and deprotonated ions were removed, molecular configurations were standardized, and duplicates were removed using the median as the averaging criterion. Then, all activities with units expressed in some form of molar concentration (nM, uM, etc.) were converted to their equivalent negative logarithm of moles per liter (pIC₅₀, pEC₅₀, etc.).

Classification threshold selection and fingerprint calculation

For each of the previously selected datasets, an active/inactive threshold was established using a binary search algorithm, with the aim of keeping the classes as balanced as possible (50%). Datasets that did not have at least five active class molecules and where a viable threshold could not be found were removed. After the data was properly filtered and cleaned, for the remaining molecules, ECFP4 descriptors (1024 bits) were calculated [24], and were used as features in posterior modelling steps. All procedures described above were done with Python scripts using the chemoinformatics library RDKit v2022.03.3 [25].

Training, Hyperparameter tuning and model validation

Classifier QSAR models were trained for each of the datasets that were obtained from the previous cleaning and filtering steps. The Random Forest (RF) algorithm [26] was used to build all the models in the AIA system. Ideally, various machine learning algorithms and hyperparameter combinations would be tested on individual datasets to identify the best-performing models, as no single algorithm is universally superior across all datasets, according to the "No Free Lunch" theorem [27]. However, in practice, time and computational constraints must be considered. The RF algorithm is often a strong choice due to its robustness and resistance to overfitting [28]. To build the models, we used scripts written in the Python programming language using the scikit-learn v1.1.1 machine learning library [29]. With the purpose of selecting only models that were statistically significant, we employed randomization tests [30] using five-fold cross validation with the Matthews Correlation Coefficient (MCC) as the evaluation metric, and selected only the models that had a p-value < 0.05. Succinctly, the randomization tests work by training and evaluating multiple models with random permutations of the same dataset and then comparing their MCCs results with the MCC of the model that was trained without the randomization procedure. The fraction of times that these permuted

MCCs are equal to or higher to the non-permuted MCC represents the p-value for this test. In order to further increase model performance, we applied hyperparameter tuning on the statistically significant models using nested cross validation with five folds as our validation/evaluation procedure. The hyperparameter chosen for adjustment was the number of trees and the hyperparameter adjustment technique was Bayesian optimization using Tree-structured Parzen Estimators (TPE). The Optuna library v3.2.0 [31] was used for the implementation of the previously described methodology. The metrics calculated with the nested cross-validation procedure were: Accuracy, precision, sensitivity, f-score, specificity and the Matthews Correlation Coefficient (MCC). Finally, in order to ensure predictive capacity, we only selected models that achieved an MCC of at least 0.5. Figure 1 provides a summarized version of the entire modeling process.

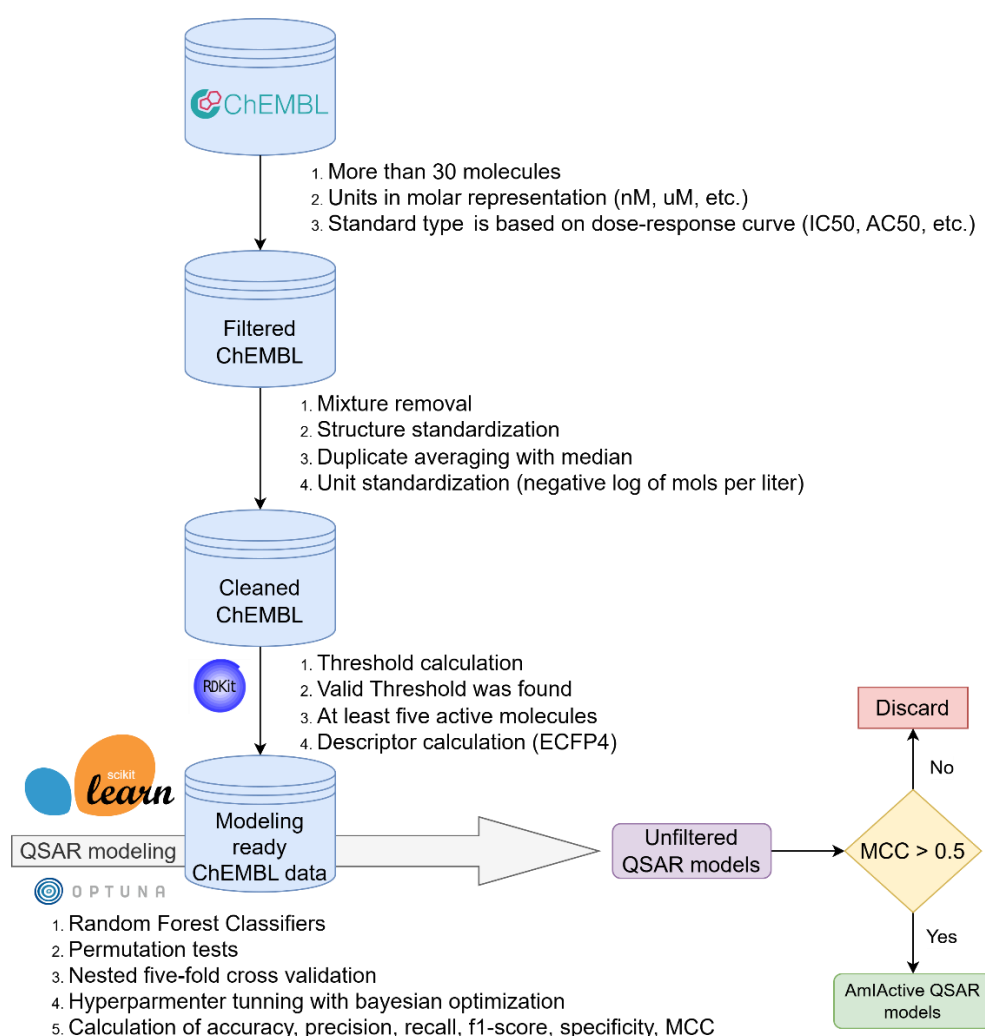


Figure 1: Detailed Schematization of the modeling process of the AIA system.

Applicability domain calculation

As important as model validation for the validity of the models predictions is the concept of applicability domain [32]. The applicability domain of a QSAR model represents the portion of chemical space in which the model predictions are deemed to be reliable given the chemical space diversity of the data that was used to train the QSAR model. In other words, if a molecule is highly different from any of the molecules present on the training data of a given model, the validation metrics of the model might not hold for this specific prediction. Therefore, if a molecule is considered to be out of the applicability domain of a given model the prediction of the model on said molecule is deemed unreliable irregardless of the predicted outcome and its probability.

The applicability domain of our models was calculated using the Python library PyADA [33] using the target datasets obtained in the steps previously described. The PyADA library uses a normalized Euclidean distance as a dissimilarity metric to quantify the distance/similarity across the different molecules. A query molecule is deemed to be inside the applicability domain if the similarity of its closest molecule on the training data was above 0.975.

Results and discussion

QSAR Models

After all the filtering and cleaning procedures, we obtained 3239 QSAR models out of the initially selected datasets. Of these 3239 models, 2277 corresponded to distinct targets of interest. The most common target type of the obtained models was single proteins representing approximately 69% of all models, followed by cell lines (15%), organisms (6%), protein complexes (4%), protein families (4%) and tissues (0.6%). Threshold and validation metrics of the QSAR models are summarized in Table 1 for different target types. Non-molecular targets (cell-lines, organisms, tissues) have lower average thresholds than their molecular counterparts. To assess the consistency of the AIA system with existing methods, we compared its predictions with those of PLATO, SwissTargetPrediction, and PASS on four targets (Supplementary Figure 1). For each target, we selected 10 molecules (5 active, 5 inactive) from the ChEMBL database, choosing among the most active and inactive tested molecules. The results indicate that while individual system performance varies by target, the overall predictions are consistent across systems.

Table 1: Average thresholds and validation metrics of the QSAR models grouped by target type.

	Threshold	Accuracy	F1-score	MCC	Precision	Recall	Specificity
Cell-line	5.40	0.82	0.81	0.65	0.84	0.81	0.83
Organism	5.47	0.82	0.81	0.65	0.83	0.81	0.82
Protein Complex	6.64	0.82	0.82	0.65	0.82	0.83	0.80
Protein Family	6.25	0.82	0.82	0.65	0.83	0.83	0.81
Single Protein	6.43	0.82	0.82	0.65	0.83	0.83	0.81
Tissue	5.49	0.82	0.82	0.65	0.83	0.82	0.82

Web utility

The AmIActive (AIA) web tool can be accessed at: amiactive.ccen.upfb.br. Upon clicking the link, the user will be directed to the AIA landing page where general information about the system and research group can be found. By clicking the “use the tool” button the user will be redirected to the graphical user interface of the actual AIA system (Figure 2).

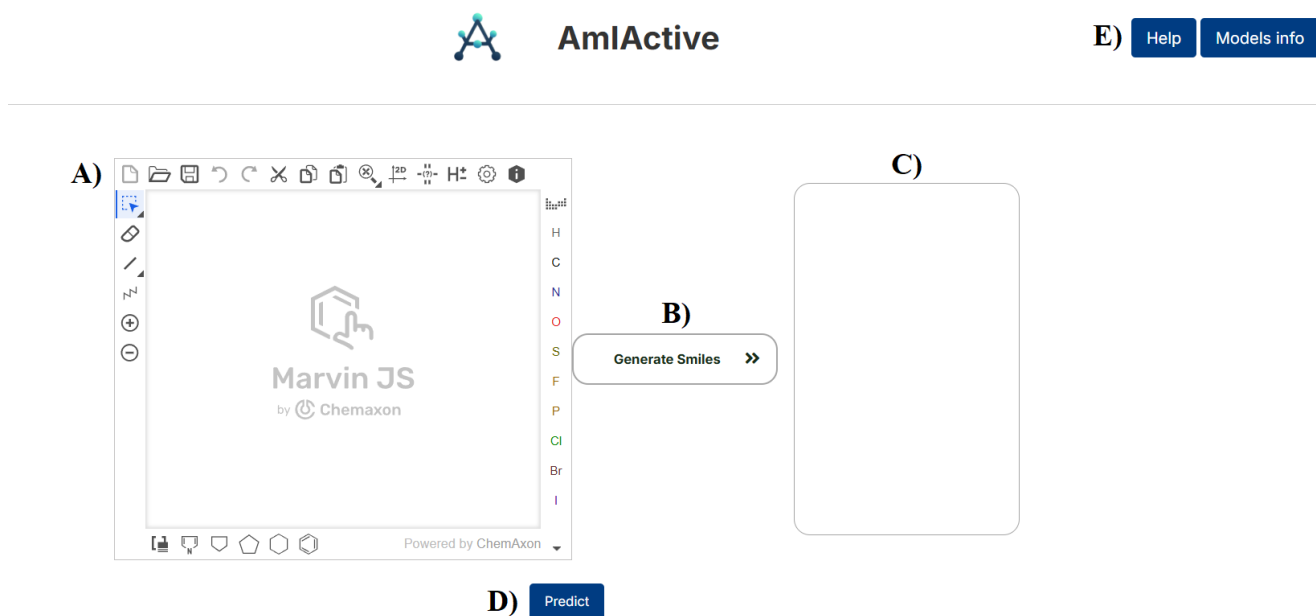


Figure 2: The AIA system user interface; A) The Marvin JS applet; B) The generate smiles button; C) The smiles text box; D) The predict button; E) The help and models info buttons.

This interface consists of the MarvinJS applet provided by Chemaxon [34] (Figure 2, A), the “generate smiles” button (Figure 2, B), a text box where the generated smiles strings are kept (Figure 2, C), the “predict” button (Figure 2, D) and the “model info” and “help” buttons (Figure 2, E) located in the upper right corner of the navigation bar. With MarvinJS, the user can draw molecular structures of interest and by clicking the generate smiles button, the drawn molecular structures will be converted to smiles strings inside the text box. Alternatively, the user may input the smiles strings directly in text box.

When ready, the user can press the “predict” button and the smiles strings present in text box will be sent to the server starting the prediction request. At this stage a box titled “Processing queue” containing the buttons “Abort prediction”, “Export to csv” and “New prediction” will appear on the screen. At this phase, the smiles strings sent to the server are already being processed or are queued for later processing when resources are available. If the user molecule is still queued for later processing, the user may terminate the prediction request by clicking the abort button, else, the user will need to wait for the current prediction request to finish before starting a new one. When the predictions are finished, the user can download the prediction results by clicking the export to csv button or start a new prediction by clicking the new prediction button.

The prediction results are outputted in a csv file, where each row represents a prediction for a user supplied molecule. This csv file has the following columns: “target”, the chembl code of model target; “standard type”, the activity type (IC50, Ki, etc.); “smiles”, the smiles string supplied by the user; “prob_inactive”, the probability that the molecule is inactive; “prob_active”, the probability that the molecule is active; “reliability”, the inclusion (reliable) or exclusion (unreliable) of the molecule in the model applicability domain; “threshold”, the cutoff point for active/inactive labeling of the molecules expressed in negative logarithm of mols per liter units; “MCC”, the model MCC; “precision”; the model precision.

Finally, if the user wants to see information about the models without any predictions present, they may click the “Models info” button. This will download a csv file containing the same information present in the prediction results file in addition to the target common name, the target type (single protein, organism, etc.), the target organism, the accuracy, the f1-score, the recall and the specificity of the model.

Future developments

To ensure the long-term reliability and effectiveness of the AIA system, we are committed to ongoing maintenance, including regular updates to address emerging user needs and resolve potential issues. Additionally, we plan to implement interfaces for individual model prediction analysis, incorporating all currently available information alongside model-agnostic explainable AI techniques such as SHAP [35], LIME [36], etc. Furthermore, we aim to enhance the system's performance by testing models trained with various machine learning algorithms—such as Gradient Boosting, Support Vector Machines, and Neural Networks—and exploring different hyperparameter combinations.

Conclusion

In this work we present the AmIActive (AIA) system, an accessible QSAR based web tool for target fishing. The AIA system currently has 3,239 models which correspond to 2,277 distinct targets, and it uses biological activity data from the ChEMBL database. The system currently holds models for single proteins, protein families, protein complexes, cell-lines, organisms and tissues. The QSAR models of the system are Random Forest classifiers trained with ECFP4 fingerprints with clearly defined applicability domains. The system is very straightforward, with a friendly user interface and the results can be downloaded in a .csv file. The AIA system is freely available to the scientific community at <https://amiactive.ccen.ufpb.br/>.

Funding sources

This work has been funded by National Council for Scientific and Technological Development (CNPq), grant number 302469/2022-2.

Acknowledgements

We would like to thank the financial support provided by CNPq and CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education).

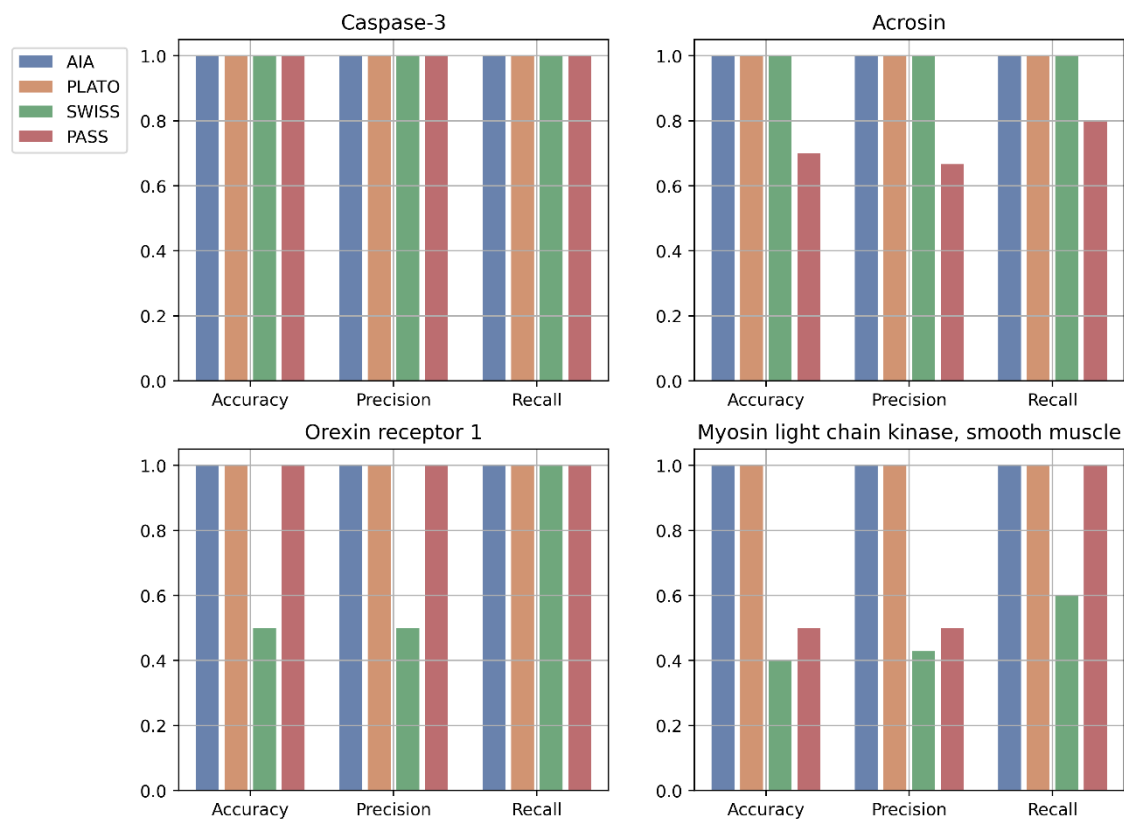
References

- [1] Macalino, S. J. Y., Gosu, V., Hong, S., Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharmacol Res.* 38 (2015), 1686-1701. doi.org/10.1007/s12272-015-0640-5
- [2] Mohammad, H. S., Khurshid, A., Sudeep, R., Jalaluddin, M. A., Mohd, A., Mohammad, H. S., ... & Inho, C. Computer aided drug design: success and limitations. *Curr. Pharm. Des.* 22.5 (2016), 572-581. doi.org/ 10.2174/1381612822666151125000550
- [3] Abramson J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630.8016 (2024), 493-500. doi.org/10.1038/s41586-024-07487-w
- [4] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic. Acids Res.* 50.D1 (2022), D439-D444. doi.org/10.1093/nar/gkab1061
- [5] AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. *BMC bioinform.* 20 (2019), 1-10. doi.org/10.1186/s12859-019-2932-0
- [6] Leman, J, K., Künze, G. Recent advances in NMR protein structure prediction with ROSETTA. *Int. J. Mol. Sci.* 24.9 (2023), 7835. doi.org/10.3390/ijms24097835
- [7] Karelina, M., Noh, J. J., Dror, R. O. How accurately can one predict drug binding modes using AlphaFold models? *Elife* 12 (2023), RP89386. doi.org/10.7554/eLife.89386.1
- [8] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Tropsha, A. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* 57(12) (2014), 4977-5010. doi.org/10.1021/jm4004285
- [9] Vázquez, J., López, M., Gibert, E., Herrero, E., & Luque, F. J. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules*, 25(20) (2020), 4723. doi.org/10.3390/molecules25204723
- [10] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J.F. Mosquera, P. Mutowo, and M. Nowotka, ChEMBL: towards direct deposition of bioassay data. *Nucleic. Acids. Res.* 47 (2019), pp. D930-D940. doi:10.1093/nar/gky1075
- [11] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J, BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic. Acids. Res.* 44(D1) (2016), D1045-D1053. doi.org/10.1093/nar/gkv1072
- [12] Venkatraman, V. FP-MAP: an extensive library of fingerprint-based molecular activity prediction tools. *Front. Chem.* 11 (2023), 1239467. doi.org/10.3389/fchem.2023.1239467

- [13] Poroikov, V. V., Filimonov, D. A., Glorizova, T. A., Lagunin, A. A., Druzhilovskiy, D. S., Rudik, A. V., ... & Pogodin, P. V. Computer-aided prediction of biological activity spectra for organic compounds: The possibilities and limitations. *Russ. Chem. Bull.*, 68 (2019), 2143-2154. doi.org/10.1007/s11172-019-2683-0
- [14] Pogodin, P. V., Lagunin, A. A., Filimonov, D. A., & Poroikov, V. V. PASS Targets: Ligand-based multi-target computational system based on a public data and naïve Bayes approach. *SAR QSAR Environ. Res.* 26(10) (2015), 783-793. doi.org/10.1080/1062936X.2015.1078407
- [15] Ciriaco, F., Gambacorta, N., Trisciuzzi, D., & Nicolotti, O. PLATO: A predictive drug discovery web platform for efficient target fishing and bioactivity profiling of small molecules. *Int. J. Mol. Sci.* 23(9) (2022), 5245. doi.org/10.3390/ijms23095245
- [16] Daina, A., Michielin, O., & Zoete, V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic. Acids. Res.* 47(W1) (2019), W357-W364. doi.org/10.1093/nar/gkz382
- [17] Zakharov, A. V., Zhao, T., Nguyen, D., Peryea, T., Shelis, T., Yasgar, A., ... & Simeonov, A. Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models. *J. Chem. Inf. Model.* 59.11 (2019), 4613-4624. doi.org/10.1021/acs.jcim.9b00526
- [18] Tullius M. S., Herrera-Acevedo, C., Menezes, R. P. B., Martin, H., Muratov, E. N., Silva, A. I. S., ... & Scotti, L. MolPredictX: online biological activity predictions by machine learning models. *Mol. Inform.* 41.12 (2022), 2200133. doi.org/10.1002/minf.202200133
- [19] Sánchez-Rodríguez, A., Pérez-Castillo, Y., Schürer, S. C., Nicolotti, O., Mangiatordi, G. F., Borges, F., ... & Cruz-Monteagudo, M., From flamingo dance to (desirable) drug discovery: a nature-inspired approach. *Drug Discov. Today* 22.10 (2017): 1489-1502. doi.org/10.1016/j.drudis.2017.05.008
- [20] Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow Jr, R. A., ... & Schneider, G. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov* 19.5 (2020), 353-364. doi.org/10.1038/s41573-019-0050-3
- [21] Iftkhar, S., de Sá, A. G. C., Velloso, J. P. L., Aljarf, R., Pires, D. E. V., Ascher, D. B. cardioToxCsM: a web server for predicting cardiotoxicity of small molecules. *J. Chem. Inf. Model* 62.20 (2022), 4827-4836. doi.org/10.1021/acs.jcim.2c00822
- [22] Daina, A., Michielin, O., Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* 7.1 (2017), 42717. doi.org/10.1038/srep42717 (2017)
- [23] 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.* 50(7) (2010), 1189. doi: 10.1021/ci100176x

- [24] Rogers, D., & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50(5) (2010), 742-754. doi.org/10.1021/ci100050t
- [25] G. Landrum, RDKit. software available at: <https://www.rdkit.org/>
- [26] Breiman, L. Random forests. *Mach. Learn.* 45 (2001), 5-32.
- [27] David, G., Rojas, A. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. *Neural Comput.* 28.1 (2016), 216-228. doi.org/10.1162/NECO_a_00793
- [28] Schonlau, M., & Zou, R. Y. The random forest algorithm for statistical learning. *Stata J.* 20(1) (2020), 3-29. doi.org/10.1177/1536867X20909688
- [29] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825-2830.
- [30] Ojala, M., & Garriga, G. C. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* 11(6) (2010).
- [31] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M.. Optuna: A next-generation hyperparameter optimization framework. *KDD '19.* (2019) (pp. 2623-2631). doi.org/10.1145/3292500.3330701
- [32] Hanser, T., Barber, C., Marchaland, J. F., & Werner, S. Applicability domain: towards a more formal definition. *SAR QSAR Environ. Res.* 27(11) (2016), 865-881. <https://doi.org/10.1080/1062936X.2016.1250229>
- [33] Dias-Silva, J. R., Oliveira, V. M., Sanches-Neto, F. O., Wilhelms, R. Z., & Júnior, L. H. Q. SpectraFP: A new spectra-based descriptor to aid in cheminformatics, molecular characterization and search algorithm applications. *Phys Chem Chem Phys.* 25(27) (2023), 18038-18047.
- [34] Chemaxon. <https://chemaxon.com/>, 2024 (accessed 27 November 2024).
- [35] Sergiu, H. Shapley value. *Game theory.* London: Palgrave Macmillan UK, 1989. 210-216. doi.org/10.1007/978-1-349-20181-5_25
- [36] Ribeiro, M. T., Singh, S. Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min..* 2016. doi.org/10.48550/arXiv.1602.04938

Supplementary Materials



Supplementary Figure 1: Comparison of the AIA, PLATO, SwissTargetPrediction and PASS systems performance on four targets.

CONCLUSÕES

Neste trabalho foi apresentado o sistema AmIActive (AIA), uma ferramenta web acessível, baseada em modelos de QSAR, desenvolvida com o objetivo de realizar *target fishing*. Atualmente, o sistema AIA conta com 3.239 modelos, correspondentes a 2.277 alvos biológicos distintos, construídos a partir de dados de atividade biológica provenientes da base de dados ChEMBL.

O sistema contempla modelos voltados para diferentes níveis de organização biológica, incluindo proteínas individuais, famílias de proteínas, complexos proteicos, linhagens celulares, organismos e tecidos. Os modelos QSAR implementados utilizam classificadores Random Forest, treinados com fingerprints ECFP4, e possuem domínios de aplicabilidade claramente definidos, assegurando maior confiabilidade nas previsões geradas.

O AmIActive foi desenvolvido com foco na usabilidade, apresentando uma interface amigável e intuitiva, além de permitir a exportação dos resultados em formato .csv, facilitando sua integração a diferentes fluxos de trabalho em pesquisa. Dessa forma, o sistema se consolida como uma ferramenta robusta e de fácil acesso, contribuindo para o avanço de estudos em quimioinformática, descoberta de fármacos e reposicionamento de compostos. O sistema AIA encontra-se disponível gratuitamente para a comunidade científica no endereço: <https://amiactive.ccen.ufpb.br/>.

ANEXOS I

Produção Científica

1. Artigo: Development of a standardized methodology for transfer learning with QSAR models: a purely data-driven approach for source task selection

Artigo aceito na revista ***SAR AND QSAR IN ENVIRONMENTAL RESEARCH***

ISSN: 1029-046X

Fator de impacto: 2.4

DOI: <https://doi.org/10.1080/1062936X.2024.2311693>

Development of a standardized methodology for transfer learning with QSAR models: a purely data-driven approach for source task selection

L. Melo, L. Scotti & M.T. Scotti

To cite this article: L. Melo, L. Scotti & M.T. Scotti (05 Feb 2024): Development of a standardized methodology for transfer learning with QSAR models: a purely data-driven approach for source task selection, SAR and QSAR in Environmental Research, DOI: [10.1080/1062936X.2024.2311693](https://doi.org/10.1080/1062936X.2024.2311693)

To link to this article: <https://doi.org/10.1080/1062936X.2024.2311693>



Published online: 05 Feb 2024.



Submit your article to this journal 



View related articles 



View Crossmark data 



Development of a standardized methodology for transfer learning with QSAR models: a purely data-driven approach for source task selection

L. Melo, L. Scotti and M.T. Scotti

Postgraduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, João Pessoa, Brazil

ABSTRACT

Transfer learning is a machine learning technique that works well with chemical endpoints, with several papers confirming its efficiency. Although effective, because the choice of source/assistant tasks is non-trivial, the application of this technique is severely limited by the domain knowledge of the modeller. Considering this limitation, we developed a purely data-driven approach for source task selection that abstracts the need for domain knowledge. To achieve this, we created a supervised learning setting in which transfer outcome (positive/negative) is the variable to be predicted, and a set of six transferability metrics, calculated based on information from target and source datasets, are the features for prediction. We used the ChEMBL database to generate 100,000 transfers using random pairing, and with these transfers, we trained and evaluated our transferability prediction model (TP-Model). Our TP-Model achieved a 135-fold increase in precision while achieving a sensitivity of 92%, demonstrating a clear superiority against random search. In addition, we observed that transfer learning could provide considerable performance increases when applicable, with an average Matthews Correlation Coefficient (MCC) increase of 0.19 when using a single source and an average MCC increase of 0.44 when using multiple sources.

ARTICLE HISTORY

Received 17 October 2023
Accepted 23 January 2024

KEYWORDS

QSAR; transfer learning; machine learning; virtual screening; multi-task learning; random forest

Introduction

Transfer learning is a machine learning technique that extends traditional supervised learning by using data pertaining to different tasks to improve the performance of a specific task. In transfer learning terms, the task we are interested in improving is called the target task, and the task we use to improve the performance of our target task is called the source task [1]. While transfer learning can provide considerable performance increases, this technique is based on the critical assumption that the target and source tasks are somehow related. If this assumption is incorrect, not only will we not observe a performance increase, but we might also degrade the performance of our target model. This effect is known as negative transfer [2].

Transfer learning was first introduced in the cheminformatics literature by Varnek et al. [3]. This pioneering work used multi-task neural networks and feature nets to model tissue – air partition coefficients, and they concluded that transfer learning could improve performance over traditional machine learning. Although multi-task neural networks are commonly thought to be a multi-task method only, this type of architecture can also be classified as a transfer learning method because the weights that predict the different tasks are shared across hidden network layers. Since the original publication by Varnek and colleagues, multiple studies using transfer learning with Quantitative Structure – Activity Relationship (QSAR) models have been published over the years [4–14].

Remarkably, very few of these published papers have addressed the critical issue of selecting suitable source datasets for a particular task of interest, either assuming suitability based on domain knowledge or simply training a transfer model with chosen source tasks and evaluating its performance.

The first study to address this issue directly was Girschick et al., who used the activity overlap measure metric [15]. This metric quantifies the chemical similarity between target and source datasets, with the rationale being that if the chemical space of the two datasets is similar, the odds of a successful transfer are higher. Although a step in the right direction, regrettably, as admitted by the authors, more than this metric is needed to capture the actual relatedness of the target and source datasets, which still needs to be assumed.

Another study that addressed this issue was by Xu et al. [16]. In this study, the authors investigated the performance of multi-task neural networks on the Kaggle Merck dataset challenge. They concluded that to achieve a successful transfer, significant similarity between the target and source datasets must exist, and the activities of similar molecules must be correlated. This study took a step further by using correlation as a measure of relatedness between the tasks, but as the authors mentioned, standard correlation coefficients cannot model nonlinear relationships. If the correlations across different QSAR tasks are not widely known, a data-driven methodology is needed.

Currently, the application of transfer learning with QSAR models is still dependent on domain knowledge of the task of interest, considerably limiting the usage of the technique by others and making it impractical to be used on large-scale systems with several different models. The problem of task relatedness is still a considerable obstacle for the large-scale application of transfer learning techniques with QSAR models [17].

In this study, we developed a transferability prediction model (TP-Model) capable of predicting the outcome of a transfer learning QSAR model (positive/negative transfer) based on transferability metrics calculated from target and source datasets. To develop this TP-Model, we first generated standard QSAR models (S-QSAR models), which were not built using transfer learning, using 14,115 datasets from the ChEMBL database [18]. In sequence, we generated 100,000 random target – source pair combinations using the same ChEMBL datasets, and for each combination, we developed a transfer learning QSAR model (T-QSAR model). By comparing the S-QSAR models and T-QSAR models results, we could identify which transfers were successful (the T-QSAR model had superior performance to the S-QSAR model). Subsequently, for each target – source pair, we calculated a set of six transferability metrics (metrics capable of separating positive transfers from negative ones) based on the target and source datasets of each combination. Finally, we trained and evaluated our TP-Model on the same random combinations using the six transferability metrics as features and the transfer outcome (positive/negative) as the

families, protein complexes, or tissues. This subgroup of targets covers more than 90% of the database and encompasses the most common cases of medicinal chemistry application.

Data cleaning, pre-processing, and descriptor calculation

Datasets with less than 30 molecules overall and units that were not percentages or could not be converted to some form of molar representation (nM, μ M, etc.) were excluded from the study. The remaining datasets were subsequently cleaned using established best practices [19]. For each dataset, we removed mixtures, deprotonated ions, standardized molecular configurations, and removed duplicates using the median as the average criterion. Afterward, we converted all the activities with unities expressed in some form of molar concentration to their equivalent form of the negative logarithm of moles per litre (pIC₅₀, pMIC, etc.). Finally, we established an active/inactive threshold for each dataset using a binary search algorithm to keep the targets as balanced as possible. Data sets that did not have at least five active class molecules and where a viable threshold could not be found were excluded. After adequately filtering and cleaning the data, we calculated ECFP4 fingerprints (1024 bits) [20] for the remaining molecules to be used as descriptors in subsequent modelling steps. All the above procedures used in-house Python scripts with its cheminformatics library RDKit [21].

Standard QSAR model (S-QSAR model) generation

To establish a baseline for performance comparison, we generated S-QSAR models for each cleaned dataset using a random forest classifier [22] with 100 decision trees. Performance was measured using repeated stratified k-fold cross-validation, with five folds and 10 repeats. To reproduce the exact train/test configuration of the cross-validation folds, the previous cross-validation procedure was performed using a random seed of 40. This step is critical to ensure that later comparisons with the equivalent T-QSAR models are performed precisely on the same data. The Matthews Correlation Coefficient (MCC) was used as the metric for performance measurement and for comparison with the equivalent T-QSAR models in later steps. The steps described above were executed with Python in-house scripts using its machine learning library scikit-learn [23].

Dataset training/validation splitting

The cleaned data was split into training and validation subgroups of equal sizes. The training split was used to generate random transfers for the TP-Model training. In contrast, the validation split was used to create the random transfers for the performance measurement. This step is necessary to prevent data leakages and ensure that the observed performance matches well with reality.

Transfer data generation

Transfer learning QSAR models (T-QSAR models)

Transfer learning and domain adaptation are considerably old and broad study subjects in the machine learning community [1,24,25]. Consequently, several transfer learning

algorithms have been developed over the years, catering to the specific needs of the relevant learning context. In this study, we have used the feature augmentation algorithm (FA) as the transfer learning algorithm for the T-QSAR models. The FA algorithm works by augmenting the feature space with three copies of the original feature space, the first being a shared feature space where both target and source features are reproduced, the second being the source only, and the last being the target only. By augmenting the feature space in this manner, a standard machine learning algorithm can learn which features help predict source-only and target-only instances and which features are shared across the two domains [26]. Although there are more modern and potentially better-performing transfer learning algorithms, such as neural network based techniques, the FA algorithm is a simple and elegant approach to the transfer learning problem. Thus, the results of this technique are expected to generalize well to more modern and powerful transfer learning methods. Another reason for choosing FA as the transfer algorithm is that it can be used on top of any machine learning algorithm, thus increasing the flexibility of the comparison with baseline S-QSAR models in subsequent steps. The described transfer model was implemented using Python's domain adaptation library ADAPT [27] using a random forest classifier as the standard machine learning method with 100 decision trees.

Custom statistical significance test

When comparing two machine learning models for the same data, one cannot simply conclude that one model is superior to another based on a marginal increase in performance, as the observed results could be due to random chance. Therefore, some type of significance testing should be performed on the observed performance differences to obtain meaningful results.

An obvious strategy would be to use a standard paired t -test on each cross-validation fold of the competing techniques. Regrettably, this approach will result in p -values that are overconfident in terms of the observed statistical significance. The reason for this is due to the fact that the cross-validation folds are not truly independent of each other [28], as each training fold shares approximately 80% of its data (in the case of 5-fold cross-validation), which violates the critical assumption of independence present in the standard paired t -test.

Some approaches that cater to this specific problem have been published in the literature [28–30] but we can create a custom significance test due to the unique nature of transfer learning (Figure 2). The custom significance test consists of a paired permutation test between a T-QSAR model and an equivalent S-QSAR baseline model and another paired permutation test between the T-QSAR model and a different transfer learning QSAR model. This model is identical to the T-QSAR model with the exception that it is built with a label permuted source dataset, which is called the permuted transfer learning QSAR model (PT-QSAR model). By permuting the source dataset labels in the context of transfer learning, we are destroying the domain information that the transfer algorithm would use to increase the performance of the target task. Thus, any performance increase compared to the baseline S-QSAR model is guaranteed to be due to random chance. Therefore, when we test the T-QSAR model against its PT-QSAR model counterpart, we are asking the question: 'What is the likelihood that the observed performance increase was due to random chance?'

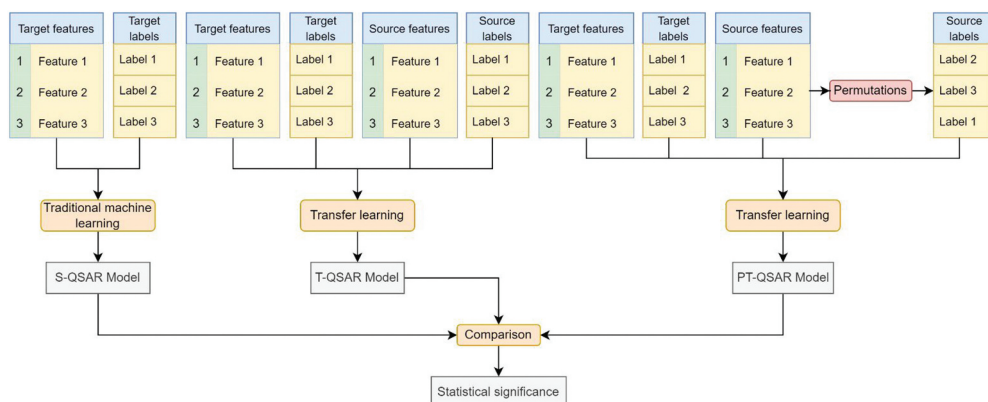


Figure 2. Detailed schematic of the custom significance test for the transfer learning models.

Random T-QSAR model generation and statistical significance calculation

To generate the transfers for analysis for both the training and validation sets, we selected 50,000 randomly chosen target – source pairs. We built a T-QSAR model for each target – source pair and evaluated it using repeated stratified k-fold cross-validation with five splits and 10 repeats with the random seed set to 40. Subsequently, using the same randomly generated target – source pairs and the same cross-validation configurations for each target – source pair, we built a PT-QSAR model with different label permutations across each fold. Finally, we calculated the T-QSAR model's p -values using our custom significance test with the cross-validation fold MCCs of the T-QSAR models, their PT-QSAR counterparts, and their equivalent baseline S-QSAR models (Figure 3).

Multiple hypothesis testing via p -value calibration

When multiple statistical significance tests are performed, it is essential to calibrate the p -values of the individual tests to adjust the desired statistical significance to the joint hypothesis (The hypothesis that all of the individual tests are significant.). Popular approaches for such objectives include Bonferroni correction and the Holm – Bonferroni method. Still, due to the unique nature of the transfer procedure outlined above, a custom multiple-hypothesis correction procedure can be developed.

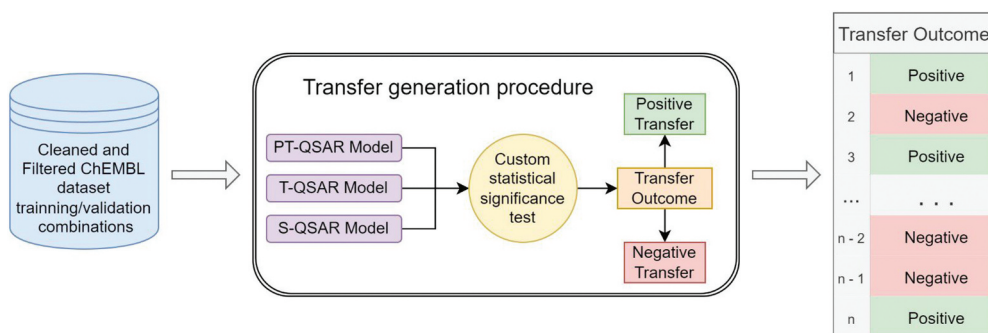


Figure 3. Transfer generation procedure schematic, from random training/validation combinations to transfer outcome ascertainment.

The custom correction procedure is achieved by repeating the random transfer generation procedure on the training set, as described above. These had the same configurations except that, instead of using the equivalent T-QSAR, S-QSAR, and PT-QSAR models, we used another PT-QSAR model in place of the T-QSAR model. By repeating the procedure with these changes and observing the proportion of statistically significant transfers with different p -value thresholds in this calibration set, we can control for the number of expected type 1 errors (rejecting a true null hypothesis).

Therefore, when applying this procedure, we are asking the question ‘What is the likelihood that an insignificant transfer will be considered significant based on the custom significance test?’ Thus, if any significant transfer is observed, it is due to random chance. By observing how many significant transfers appear for a given p -value threshold in this calibration set, we can estimate the number of false positives (transfers falsely concluded as significant) in the actual data.

Calculation of transferability metrics

By leveraging relevant domain information, transfer learning can significantly increase the performance of a target task. However, the critical assumption of domain similarity must be satisfied for this to be true. Therefore, if we can quantify the domain similarity between two tasks, in theory, we could reliably predict the outcome of a transfer before the transfer is made. With this in mind, we used six metrics to quantify the transferability potential between the target and source tasks. The transferability metrics, described in the following sections, were applied to the training and validation sets using custom in-house Python scripts.

Number of overlapping molecules

One of the relevant similarity factors across different tasks is the proximity of individual instances based on the feature space. In a QSAR context, this could be phrased as the overall molecular similarity between the target and source datasets. Considering this, one way to quantify such information would be computing the number of molecules in both the target and source datasets. Thus, by comparing the smile strings of the target dataset and its potential source, we can calculate the number of overlapping molecules for both train and validation target – source pairs.

Average closest target – source molecule similarity

Molecules can be significantly similar without being necessarily equal. Therefore, simply considering the molecules that are present in both the target and source data will fail to consider the cases where a low overlap but considerable molecular similarity is present. One solution to this problem is to calculate all the similarities between the target data molecules and the source ones, select the most similar target – source pairs for each target molecule, and calculate the mean of their similarities. Thus, a low value will indicate that, on average, for each molecule on the target data, at least one considerably similar molecule exists in the source data. This metric was implemented using the Manhattan distance with ECFP4 fingerprints [20] with a bit size of 1024.

Domain classifier

Another approach used to quantify the descriptor space similarity between datasets is the domain classifier [31]. The domain classifier is a model trained to differentiate source dataset instances from target ones. If the classifier performs well, target data are easily discernible from source data, indicating a low overall similarity between the datasets. If the opposite is true, the datasets are considerably similar.

Source model score permutation test

Although important, the similarity of the target and source datasets is not the only factor determining transfer outcomes. Even if the two datasets are considerably similar, the relationship between the features and the variable of interest might differ considerably. In a QSAR setting, certain structural aspects of a molecule might influence the outcome of their respective targets differently. The SAR of the target and source tasks must be similar to achieve a successful transfer. For this purpose, we developed a metric capable of evaluating the closeness between the source and target SAR using individual source QSAR models. This metric is based on the idea that if a model trained with source data is also capable of predicting the target task with some reliability, then the underlying SAR of the source must have some degree of similarity with the target SAR. The implementation of this approach is as follows. First, we made predictions on the target data using a QSAR model that was trained only with source data. In sequence, we calculated the MCC of the source model for the target data using the source model predictions and the actual target labels. Subsequently, we generated 100 different permutations of the source model predictions for the target data and calculated the MCC for each of these permutations. Finally, we compared the untouched MCC with its permutations, calculating the proportion of times a permutation MCC was equal to or greater than the untouched one. This procedure is a statistical significance test that estimates the likelihood that the observed source model MCC on the target data was due to random chance. This metric will be close to zero if the source task has a similar SAR to the target task.

Consistent biological assays

Another potential source of information, unrelated to the chemical space or the SAR of the particular tasks, is the type of biological assay used to determine the molecular activities (IC₅₀, K_i, etc.). Although the chemical space information and the SAR of the targets are technically invariant concerning the biological assay used to determine the individual molecular activities, transfers with source tasks that were tested using the same test as their target counterparts are expected to be more successful.

Consistent units

Following the same logic described above, transfers whose outcomes were measured using the same units are expected to have a higher likelihood of success than transfers involving targets and sources with different units.

Transferability prediction model (TP-Model)

Using the *p*-value threshold obtained from the calibrated transfer *p*-values calculated in previous steps, we classified the transfer outcomes to both the train and validation sets as

positive or negative, depending on the p -value of the transfer. Then, we merged the transferability metrics matrices of both train and validation sets with their corresponding transfer outcomes, creating a supervised machine learning classification problem where the features are the transferability metrics and the transfer outcome is the variable of interest. Owing to the extreme imbalance in the data, we used a balanced random forest classifier [32] as our TP-model algorithm by using Python's imbalanced machine learning library imbalanced-learn [33]. Finally, we trained the TP-model with 1000 decision trees on the training set and evaluated its performance on the validation set.

TP-Model applicability and validation

To evaluate the TP-Model in an actual application scenario, we randomly selected 100 datasets from the validation set subjected to biological assay and unit restrictions. These restrictions are necessary because, in a real-world scenario, a researcher will most likely not model datasets with certain types of biological assays and datasets with non-standard units. For each of the selected datasets, we calculated their transferability metrics with all the other available datasets and, subsequently, predicted the transfer outcome of all the combinations using the created TP-Model. Then, we built T-QSAR and PT-QSAR models for combinations predicted as positive by the TP-Model. We evaluated the outcome of the obtained transfers using the developed custom significance test and the calibrated p -value threshold. If more than one positive transfer was observed for the same target, we applied multi-source (MS) transfer learning using a maximum of 15 significant source datasets, ranked by performance increase.

Results and discussion

p -value calibration

By progressively decreasing the p -value threshold for the multiple test calibration set, we evaluated the number of false positive transfers expected for a custom significance test p -value. Our results indicate that conclusively separating significant transfers from insignificant ones in a multiple-test scenario is challenging because, even at a test p -value of 0, some false positive transfers (type 1 errors) remain. More specifically, of the 50,000 transfers on the calibration set, four were still considered significant with a p -value threshold of 0. Thus, the probability that an insignificant transfer will achieve a p -value of 0 is 0.008%. Nevertheless, setting the p -value threshold to 0 clearly shows that transfer learning can be applied to QSAR settings with proper data. More precisely, using a threshold of 0, we obtained 62 positive transfers on the training set, which is 16 times higher than expected if no significant transfers existed.

Transfer results analysis

Of the 50,000 random transfers in the training set, 62 were considered significant, whereas, of the 50,000 random transfers in the validation set, 63 were considered significant. These results indicate a remarkable consistency in the probability of finding a significant transfer in a random sample space, as even with datasets composed of

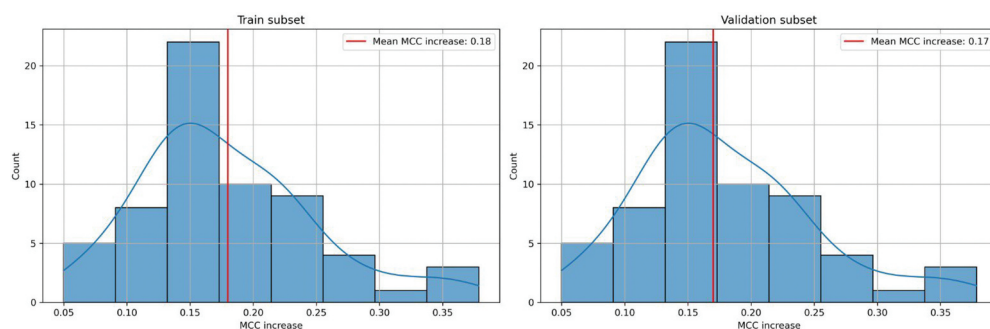


Figure 4. Mean MCC increase (difference between the transfer model MCC and baseline MCC) for the train and validation subset transfers.

distinct targets, the probability of finding a significant transfer in a random combination of targets is 0.12%. The results demonstrate that significant transfers are rare. Thus, a naïve approach to selecting appropriate source tasks is bound to be costly. Notably, though, the few significant transfers observed in the train and validation sets had considerable increases in performance, with an average MCC increase of approximately 0.17 and a maximum of 0.37 (Figure 4), demonstrating that transfer learning can increase the performance of S-QSAR models beyond marginal results.

Transferability metrics analysis

By comparing the transfer metrics patterns of the significant transfers with their insignificant counterparts, we could observe the capacity of the transfer metrics for separating the significant transfers from the insignificant ones. Based on these observations, we have determined that all transfer metrics utilized can select some significant transfers. All of the transfer metrics demonstrated predictive capabilities, and all of their observed patterns matched the intuition behind their conception (Figure 5). The number of overlapping molecules was much higher in the significant transfers, with the average number of overlapping molecules being 102, whereas insignificant transfers had an average of 0.6. The average of the mean closest molecule similarity was much lower in the significant transfers, which had an average of 9.62, whereas insignificant transfers had an average of 65.62. The domain classifier metric had an average of 0.34 for significant transfers, while insignificant transfers had an average of 0.96. As expected, the source model score permutation test had much lower average values for significant transfers, with an average of 0.05 for significant transfers and 0.68 for insignificant ones. Finally, for both the consistency of biological assays and consistency of units, significant transfers had higher averages, with averages of 0.95 and 0.97 for significant transfers and averages of 0.11 and 0.46 for insignificant ones, respectively.

Transferability prediction model (TP-Model) analysis

Our results indicate that the transferability between two datasets is not only predictable but also highly efficient. This is evidenced by the precision of 17% achieved by our TP-

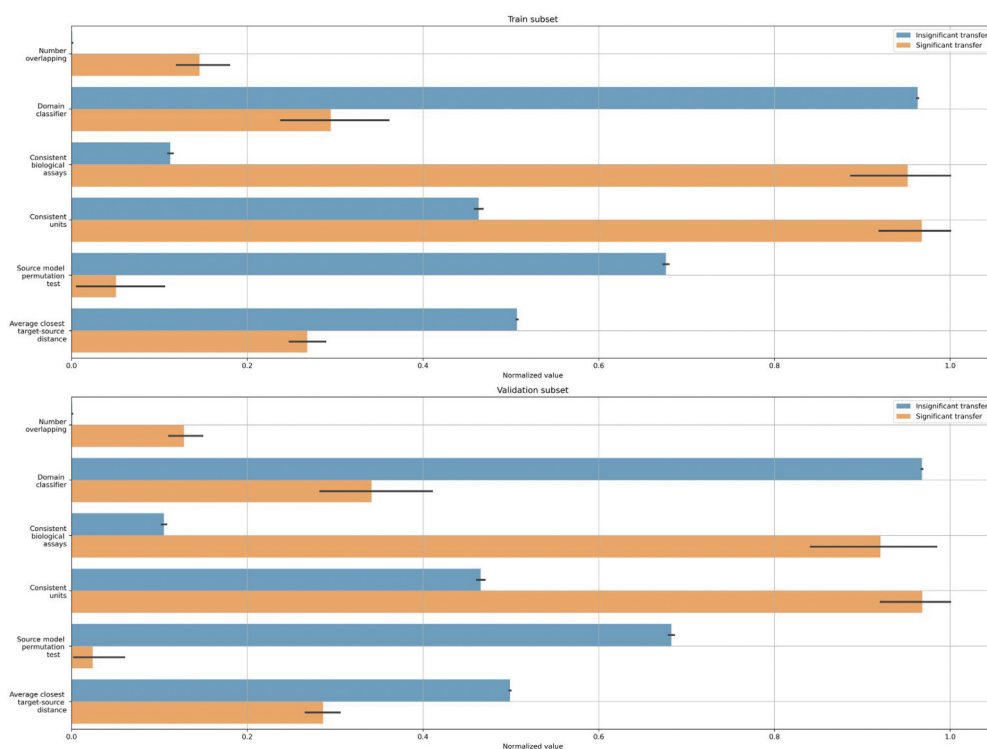


Figure 5. Normalized mean of the transferability metrics calculated based on significant transfers (orange bars) and insignificant transfers (blue bars).

Model, in contrast with the 0.126% proportion of positive transfers for the validation set. Considering the extreme imbalance in the validation subset data, this represents a considerable 135-fold increase in performance compared with a random search. Besides the considerable increase in performance, the TP-Model could also capture the vast majority of the positive transfers present, with a 92% sensitivity rate.

TP-Model applicability analysis

Of the 100 randomly selected targets from the validation set that were used to test the applicability of the TP-Model, 17 had at least one significant transfer identified (Table 1). These results demonstrate that obtaining a successful transfer is far from a guarantee, even with many targets available as potential source tasks. Nevertheless, the successful transfers targets have diverse target types, with distinct organisms, cell-lines and proteins present.

Notably, the number of significant transfers varied considerably across the different validation targets, with some validation targets achieving as many as 300 significant transfers. In contrast, others could only accomplish a single significant transfer. This variability was also present in the observed performance increase of the T-QSAR models, where the average MCC increase was 0.19. Still, some models achieved an MCC increase as high as 0.2, while others achieved an MCC increase as low as 0.06. Upon further analysis,

Table 1. Biological assays from the applicability datasets that had at least one successful transfer identified.

<i>Target code</i>	<i>Biological assay</i>	<i>Nº of successful transfers</i>	<i>Nº of transfers predicted positive</i>	<i>Nº of potential sources</i>	<i>Best transfer MCC increase</i>	<i>MS transfer MCC increase</i>
CHEMBL1075386	IC50	397	693	13474	0.255834	0.58169
CHEMBL1075489	IC50	394	695	13474	0.334446	0.60592
CHEMBL1075438	IC50	369	658	13474	0.479142	0.834312
CHEMBL2366121	IC50	298	753	13474	0.262394	0.544914
CHEMBL2366197	IC50	215	579	13474	0.360973	0.585371
CHEMBL1075509	IC50	154	751	13474	0.231924	0.433809
CHEMBL614610	LC50	69	253	13474	0.110238	0.394687
CHEMBL614300	GI50	43	316	13474	0.061183	0.268759
CHEMBL614860	GI50	40	316	13474	0.061297	0.313565
CHEMBL390	LC50	22	215	13474	0.077503	0.228442
CHEMBL4575	Ki	2	168	13474	0.081810	0.086253
CHEMBL1741176	IC50	1	71	13474	0.073578	N/A
CHEMBL1741192	AC50	1	95	13474	0.049946	N/A
CHEMBL1936	Kd	1	84	13474	0.128261	N/A
CHEMBL2068	Kd	1	66	13474	0.245387	N/A
CHEMBL5426	Kd	1	112	13474	0.307846	N/A
CHEMBL613963	IC50	1	61	13474	0.142562	N/A

we identified that this variability can be partially explained by the target type (single protein, organism, cell line, etc.).

More specifically, the results indicate that cell lines are highly transferable between each other, responsible for cases where hundreds of significant transfers are observed, and considerable performance increases were obtained. These results match well with intuition, as even though different cell lines might have different structures and properties, they share many molecular structures overall. Thus, even if two cell lines are distinct, several molecular interactions will be consistent across these different cell lines.

However, proteins tend to be much more specific structures by nature. Therefore, finding another protein with sufficient molecular similarity to produce a successful transfer is much more challenging. Nevertheless, although more limited, transfer learning can be used with protein targets. Notable transfers involving protein targets include the following: 1) the transfer between mycobacterium tuberculosis proteins replicative DNA helicase (CHEMBL1741192, AC50) and RecA (CHEMBL1741171, EC50), with an MCC increase of 0.05, which are both involved in DNA synthesis and metabolism; 2) the transfer between the human proteins stem cell growth factor receptor (CHEMBL1936, Kd) and platelet-derived growth factor receptor alpha (CHEMBL2007, Kd), with an MCC increase of 0.13, which are two different types of growth factor receptors of oncologic interest; 3) the transfer between the human proteins ephrin type-A receptor 2 (CHEMBL2068, Kd) and ephrin type-A receptor 4 (CHEMBL3988, Kd), with an MCC increase of 0.24, which are two subtypes of the same receptor; 4) the transfer between the Homo sapiens X-box-binding protein 1 (CHEMBL1741176, IC50) and the Mus musculus DNA damage-inducible transcript 3 protein (CHEMBL2146304, IC50), with an MCC increase of 0.07, which are transcription factors associated with Endoplasmic Reticulum Stress response; 5) the transfer between the human proteins Dual-specificity tyrosine-phosphorylation regulated kinase 3 (CHEMBL4575, Ki) and the cGMP-dependent protein kinase 1 beta (CHEMBL4273, Ki), with an MCC increase of 0.08, which are human kinases.

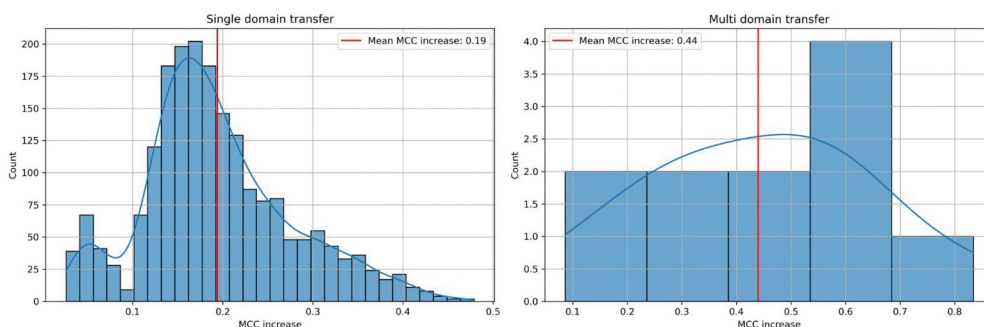


Figure 6. Mean MCC increase (difference between the transfer model MCC and baseline MCC) on the applicability data for single-source and multi-source transfers.

Of the 17 target tasks with at least one significant transfer, 11 had more than one significant transfer identified and, thus, were suitable for MS transfer learning. The MS transfer learning QSAR models demonstrated superior performance to their single-domain counterparts and, therefore, should be preferred whenever possible. The average MCC increase on these MS transfer learning QSAR models was 0.44 (Figure 6), and all models had superior performance to their best single-domain counterparts. Remarkably, the MS transfer learning QSAR model of the EFE-184 cell line with ChEMBL target code ChEMBL1075438 achieved an MCC of 0.82. In contrast, its baseline line S-QSAR model could not achieve statistical significance, with an MCC of -0.01 .

Conclusion

In this study, we examined the potential applications and limitations of transfer learning with QSAR models. Specifically, we sought to answer the following questions. ‘When should we be confident that an observed transfer is significant?’ ‘What factors are important for a successful transfer?’ ‘Which datasets should we use as sources for transfer learning with QSAR models?’ To answer these questions, we used the datasets in the ChEMBL database to generate 100,000 random transfers, and based on these transfers, we studied the overall transfer behaviour. We addressed the limitations of using significance testing with cross-validation folds by developing a custom significance test to compare the S-QSAR model with the T-QSAR model counterpart. Previous evaluations of transfer learning QSAR models in the cheminformatics literature were performed based on the difference in performance between the transfer models and their standard counterparts. Although this approach is direct and easy to apply, it does not consider the possibility that the observed performance increase is simply due to increased model complexity. Thus, using our custom significance test, researchers can be more confident in their transfer results. To comprehend successful transfer behaviour, we used six transferability metrics. Based on the significant transfers obtained in the random transfers, we could understand how the different metrics affected the transfer learning outcome. Using the proposed metrics, we developed a transferability prediction model (TP-Model) capable of predicting the

transferability between two datasets. Our applicability data results show that cell lines are much more transferable than other target types and can serve as optimal targets for transfer learning application. Although more limited, transfer learning can also be used with protein targets with several successful transfer target – source protein pairs exhibiting a clear and intuitive transferability relation. Our results indicate that, whenever feasible, MS transfer learning should be preferred. The MS models of all 11 targets with more than one successful transfer performed better than all of their single-source counterparts. With our TP-Model, a researcher can reap the potential benefits of transfer learning without previous knowledge of the tasks involved, considerably decreasing the work required to implement a successful transfer model. The TP-Model can be acquired by making a request from the author, L. Melo, via email (luisfelipe.melo@hotmail.com) or by replicating the methodological steps outlined in this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico [302469/2022-2].

ORCID

M.T. Scotti  <http://orcid.org/0000-0003-4863-8057>

References

- [1] S.J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Trans. Knowl. Data Eng. 22 (2009), pp. 1345–1359. doi:[10.1109/TKDE.2009.191](https://doi.org/10.1109/TKDE.2009.191).
- [2] M.T. Rosenstein, Z. Marx, L.P. Kaelbling, and T.G. Dietterich, *To transfer or not to transfer*, NIPS workshop on transfer learning, British Columbia, CA, 2005. Vol. 898, 3.
- [3] A. Varnek, C. Gaudin, G. Marcou, I. Baskin, A.K. Pandey, and I.V. Tetko, *Inductive transfer of knowledge: Application of multi-task learning and feature net approaches to model tissue-air partition coefficients*, J. Chem. Inf. Model. 49 (2009), pp. 133–144. doi:[10.1021/ci8002914](https://doi.org/10.1021/ci8002914).
- [4] K. Liu, X. Sun, L. Jia, J. Ma, H. Xing, J. Wu, H. Gao, Y. Sun, F. Boulnois, and J. Fan, *Chemi-Net: A molecular graph convolutional network for accurate drug property prediction*, Int. J. Mol. Sci. 20 (2019), pp. 3389. doi:[10.3390/ijms20143389](https://doi.org/10.3390/ijms20143389).
- [5] S. Kearnes, B. Goldman, and V. Pande, *Modeling industrial ADMET data with multitask networks*, arXiv preprint (2016). doi:[10.48550/arXiv.1606.08793](https://doi.org/10.48550/arXiv.1606.08793).
- [6] Z. Ye, Y. Yang, X. Li, D. Cao, and D. Ouyang, *An integrated transfer learning and multitask learning approach for pharmacokinetic parameter prediction*, Mol. Pharm. 16 (2018), pp. 533–541. doi:[10.1021/acs.molpharmaceut.8b00816](https://doi.org/10.1021/acs.molpharmaceut.8b00816).
- [7] X. Li, Z. Li, X. Wu, Z. Xiong, T. Yang, Z. Fu, X. Liu, X. Tan, F. Zhong, and X. Wan, *Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation*, J. Med. Chem. 63 (2019), pp. 8723–8737. doi:[10.1021/acs.jmedchem.9b00855](https://doi.org/10.1021/acs.jmedchem.9b00855).

- [8] E.N. Feinberg, E. Joshi, V.S. Pande, and A.C. Cheng, *Improvement in ADMET prediction with multitask deep featurization*, J. Med. Chem. 63 (2020), pp. 8835–8848. doi:10.1021/acs.jmedchem.9b02187.
- [9] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, and V. Svetnik, *Deep neural nets as a method for quantitative structure–activity relationships*, J. Chem. Inf. Model. 55 (2015), pp. 263–274. doi:10.1021/ci500747n.
- [10] G.E. Dahl, N. Jaitly, and R. Salakhutdinov, *Multi-Task neural networks for QSAR predictions*, arXiv preprint (2014). doi:10.48550/arXiv.1406.1231.
- [11] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, *Massively multitask networks for drug discovery*, arXiv preprint (2015). doi:10.48550/arXiv.1502.02072.
- [12] K. Abbasi, A. Poso, J. Ghasemi, M. Amanlou, and A. Masoudi-Nejad, *Deep transferable compound representation across domains and tasks for low data drug discovery*, J. Chem. Inf. Model. 59 (2019), pp. 4528–4539. doi:10.1021/acs.jcim.9b00626.
- [13] G.B. Goh, C. Siegel, A. Vishnu, and N. Hodas, *Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction*, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 2018. pp. 302–310. doi:10.1145/3219819.3219838.
- [14] X. Li and D. Fourches, *Inductive transfer learning for molecular activity prediction: Next-gen QSAR models with MolPMoFIT*, J. Cheminf. 12 (2020), pp. 1–15. doi:10.1186/s13321-020-00430-x.
- [15] T. Girschick, U. Rückert, and S. Kramer, *Adapted transfer of distance measures for quantitative structure-activity relationships and data-driven selection of source datasets*, Comput. J. 56 (2013), pp. 274–288. doi:10.1093/comjnl/bxs092.
- [16] Y. Xu, J. Ma, A. Liaw, R.P. Sheridan, and V. Svetnik, *Demystifying multitask deep neural networks for quantitative structure–activity relationships*, J. Chem. Inf. Model. 57 (2017), pp. 2490–2504. doi:10.1021/acs.jcim.7b00087.
- [17] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, and J. Pei, *Transfer learning for drug discovery*, J. Med. Chem. 63 (2020), pp. 8683–8694. doi:10.1021/acs.jmedchem.9b02147.
- [18] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J. F. Mosquera, P. Mutowo, and M. Nowotka, *ChEMBL: Towards direct deposition of bioassay data*, Nucleic. Acids. Res. 47 (2019), pp. D930–D940. doi:10.1093/nar/gky1075.
- [19] D. Fourches, E. Muratov, and A. Tropsha, *Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research*, J. Chem. Inf. Model. 50 (2010), pp. 1189. doi:10.1021/ci100176x.
- [20] D. Rogers and M. Hahn, *Extended-connectivity fingerprints*, J. Chem. Inf. Model. 50 (2010), pp. 742–754. doi:10.1021/ci100050t.
- [21] G. Landrum, *RDKit software*. Available at <https://www.rdkit.org/>
- [22] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32. doi:10.1023/A:1010933404324
- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, *Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res. 12 (2011), pp. 2825–2830.
- [24] K. Weiss, T.M. Khoshgoftaar, and D. Wang, *A survey of transfer learning*, J. Big Data 3 (2016), pp. 1–40. doi:10.1186/s40537-016-0043-6.
- [25] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, *A comprehensive survey on transfer learning*, Proc. IEEE 109 (2020), pp. 43–76. doi:10.1109/JPROC.2020.3004555.
- [26] H. Daumé III, *Frustratingly easy domain adaptation*, arXiv preprint (2009). doi:10.48550/arXiv.0907.1815.
- [27] A. de Mathelin, F. Deheeger, G. Richard, M. Mougeot, and N. Vayatis, *Adapt: Awesome domain adaptation python toolbox*, arXiv preprint (2021). doi:10.48550/arXiv.2107.03049.
- [28] C. Nadeau and Y. Bengio, *Inference for the generalization error*, Adv. Neur. In. 12 (1999), pp. 307–313.
- [29] R.R. Bouckaert and E. Frank, *Evaluating the replicability of significance tests for comparing learning algorithms*, Lect. Notes. Artif. Int. 8th Pacific-Asia Conference, PAKDD, Sydney, AU, 2004. doi:10.1007/b97861.

- [30] T.G. Dietterich, *Approximate statistical tests for comparing supervised classification learning algorithms*, Neural. Comput. 10 (1998), pp. 1895–1923. doi:[10.1162/089976698300017197](https://doi.org/10.1162/089976698300017197).
- [31] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, *Domain-adversarial training of neural networks*, J. Mach. Learn. Res. 17 (2016), pp. 2096–2030.
- [32] C. Chen, A. Liaw, and L. Breiman, *Using Random Forest to Learn Imbalanced Data*, Vol. 110, University of California, Berkeley, 2004. p. 24.
- [33] G. Lemaître, F. Nogueira, and C.K. Aridas, *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning*, J. Mach. Learn. Res. 18 (2017), pp. 559–563.

ANEXOS II

Eventos e Congressos

1. XXXI Symposium on Bioinformatics and Computer-Aided Drug Discovery

Título: Amiactive (Aia): A Large-Scale Qsar Based Target Fishing And Polypharmacology Predictive Web Tool

Tipo de apresentação: YSC Flash speaker/Poster

Link da apresentação:

https://way2drug.com/bcadd2025/content/video/Melo_LBCADD-2025.mp4



CERTIFICATE of e-POSTER PRESENTATION

presented for

**AMIACTIVE (AIA): A LARGE-SCALE
QSAR BASED TARGET FISHING AND
POLYPHARMACOLOGY PREDICTIVE WEB TOOL**

M. T. Scotti, L. Melo

at the

**XXXI Symposium on
Bioinformatics and Computer-Aided Drug Discovery**

2025 / October 20-22

Symposium Co-Chairs

Member of Rus. Acad. Sci.
Vladimir V. Poroikov

Professor
Roman G. Efremov



FLASH SPEAKER CERTIFICATE

presented to

Luis Felipe Melo

Federal University of Paraiba

in recognition of outstanding contribution as flash speaker of the presentation entitled

**AMIACTIVE (AIA): A LARGE-SCALE QSAR BASED TARGET FISHING
AND POLYPHARMACOLOGY PREDICTIVE WEB TOOL**

at the

**XXXI Symposium on
Bioinformatics and Computer-Aided Drug Discovery**

2025 / October 20-22

Symposium Co-Chairs

*Member of Rus. Acad. Sci.
Vladimir V. Poroikov*

*Professor
Roman G. Efremov*

DIPLOMA

XXXI Symposium on
Bioinformatics and Computer-Aided Drug Discovery

2025 / October 20 - 22

THIRD DEGREE
Luis Felipe Melo
Federal University of Paraiba

AMIACTIVE (AIA): A LARGE-SCALE QSAR BASED TARGET FISHING
AND POLYPHARMACOLOGY PREDICTIVE WEB TOOL

Symposium Co-Chairs


Member of Rus. Acad. Sci.
Vladimir V. Poroikov


Professor
Roman G. Efremov