Interpretação de Grandes Volumes de Dados Cromatográficos no Controle de Dopagem e Semântica das Substâncias com Integração no PubChem

Guy M. B. Junior[1], Sergio Manuel Serra da Cruz[1], Jorge Zavaleta[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro

Rio de Janeiro - RJ – Brasil

guyjunior@iq.ufrj.br, serra@ppgi.ufrj.br, zavaleta@pet-si.ufrrj.br

Resumo. *Este trabalho apresenta uma proposta computacional para auxiliar na interpretação de grandes volumes de dados baseada na lógica fuzzy e uso do coeficiente de determinação R^2 aplicados a dados oriundos de cromatógrafos líquidos de alta eficiência acoplados a espectrômetros de massas (HPLC-MS). A abordagem apoia analistas na detecção semiautomatizada de substâncias alvo nas amostras de urina de atletas submetidos ao controle de dopagem. Além disso, o método integra informações da plataforma PubChem, que fornece dados químicos detalhados, como estrutura molecular, propriedades físico-químicas e identificadores únicos, enriquecendo a análise e interpretação dos resultados. Os primeiros resultados indicam que o método não apenas acelera a detecção, mas também permite a identificação simultânea de múltiplas substâncias alvo, fornecendo subsídios adicionais para a caracterização de compostos e facilitando a identificação de novos padrões de dopagem.*

1. Introdução

Novas substâncias dopantes e novos métodos de dopagem em atletas são problemas que requerem constante atenção por parte de governos e organizações desportivas. Somente em 2012, os gastos com controle de dopagem aproximaram-se de US$ 500 milhões [Maennig 2014]. Logo, é urgente desenvolver novas estratégias que apoiem as atividades de detecção de doping, pois a lista de substâncias e métodos proibidos é atualizada anualmente e a sofisticação da dopagem acompanha, par e passo, a evolução da Farmacologia e da Medicina Desportiva [Aquino Neto 2001].

Embora os avanços científicos propiciem a melhora da detecção de doping, há um crescente esforço para desenvolver novas estratégias para evitar a dopagem e manipulação dos resultados esportivos [Maennig 2014]. Nesse cenário, organizações antidopagem, como a Agência Mundial Antidoping (WADA), têm um papel fundamental na luta contra a dopagem [Pereira 2022]. As agências buscam proporcionar maior integridade, garantir condições justas nas competições e preservar a saúde dos atletas.

Mundialmente, são coletadas milhares de amostras de urina para fins de controle de dopagem. O processamento analítico gera big data com terabytes de dados brutos a cada nova competição. Esses dados são altamente sensíveis e protegidos, possuindo uma grande variedade de formatos e riqueza de informações. No entanto, seu potencial ainda

não é plenamente explorado para uso em pesquisas na Farmacologia, Medicina Desportiva ou mesmo na E-Ciência [Gleaves et al. 2021].

Atualmente, o Laboratório Brasileiro de Controle de Dopagem (LBCD) é o único laboratório acreditado pela WADA na América do Sul. Ele é responsável pela análise de aproximadamente 7.000 amostras/ano, um número que cresce anualmente. Os analistas do LBCD avaliam cada amostra individualmente. Eles inspecionam os dados manualmente através de cromatogramas (sob a forma de relatório ilustrado na Figura 1) gerados por equipamentos científicos do tipo cromatógrafos líquidos de alta eficiência acoplados a espectrômetros de massas (HPLC-MS). Esse procedimento, apesar de robusto e padronizado, é caro e pode ser propenso a erros humanos devido ao grande volume de informações e ao curto tempo de análises, entre outros [Mogollon et al. 2014].

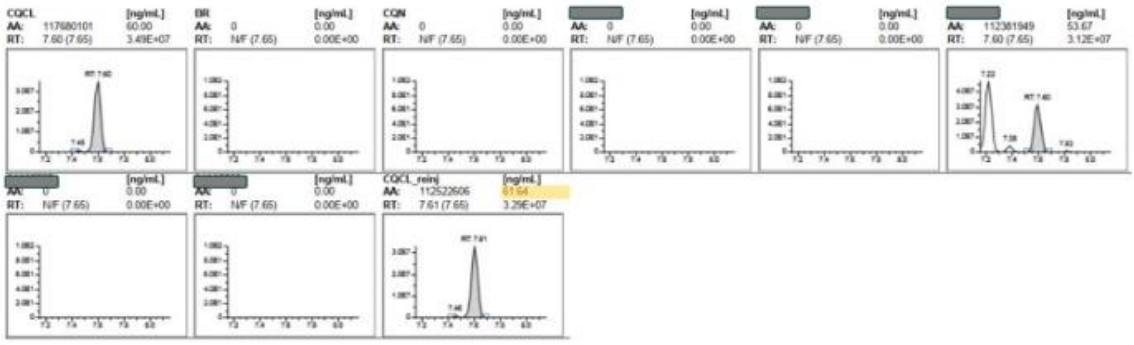

Figura 1. Exemplo de relatório extraído para os analistas interpretarem

Figura 1. Exemplo de relatório com as gaussianas de cromatogramas (picos cromatográficos) para a substância efedrina-D3 obtidos através de HPLC-MS.

Adicionalmente, para assegurar sua acreditação internacional, anualmente o LBCD passa por rigorosos testes de controle de qualidade. A WADA envia amostras fortificadas com substâncias alvo para que através de marchas analíticas os analistas e o laboratório as identifiquem e cumpram os requisitos de acreditação. Após a conclusão, o laboratório consolida os resultados reportando-os à WADA, assegurando que todas as análises estejam em conformidade com os padrões internacionais. No entanto, esses testes sobrecarregam ainda mais os analistas.

Nesse contexto, o uso de ferramentas computacionais que integrem bases de dados químicas como o PubChem surge como uma abordagem estratégica. O PubChem é uma plataforma pública mantida pelo National Center for Biotechnology Information (NCBI) que fornece informações detalhadas sobre substâncias químicas, como nomes sistemáticos, identificadores únicos (CAS), estrutura molecular e propriedades físico-químicas [Kim et al. 2016]. Essas informações podem ser usadas para enriquecer a interpretação dos resultados de cromatogramas e fornecer subsídios adicionais para a caracterização de substâncias suspeitas em análises antidopagem. Assim, ao associar dados analíticos com informações do PubChem, é possível obter um panorama mais abrangente sobre os compostos identificados, facilitando a detecção de padrões emergentes de dopagem.

Logo, com o objetivo de agilizar esses processos, defende-se a criação de um novo método computacional para apoiar os analistas a realizarem análises semiautomatizadas dos cromatogramas. O método proposto é baseado em pipelines que usam lógica fuzzy [Marro et al. 2011] e ciência de dados através do coeficiente de determinação R^2 [Chiode 2021] para apoiar a interpretação de dados dos cromatogramas. A abordagem visa auxiliar

os profissionais no processo de tomada de decisões e reduzir o tempo e as possibilidades de erros humanos. Além disso, o método fornece sugestões de possíveis interpretações para cada substância alvo presente nas amostras analisadas, sinalizando novos padrões de dopagem.

2. Trabalhos Relacionados

Essa seção aborda trabalhos relacionados à utilização de lógica fuzzy, coeficiente de determinação R^2 e bases de dados químicas em análises cromatográficas. No contexto de avaliação não linear, Ukić et al. (2022) demonstram a eficácia da lógica fuzzy no apoio à classificação de substâncias do tipo açúcares usando modelos quantitativos de relação estrutura-retenção (QSRR). Essa abordagem permite a definição de critérios flexíveis para a detecção das substâncias nos cromatogramas, melhorando a precisão da análise quando comparada a métodos booleanos.

O coeficiente de determinação R^2 é uma medida estatística que indica a proporção da variância dos dados de um modelo. Abed e Rasheed (2020) aplicaram o R^2 para avaliar a qualidade do ajuste de modelos de regressão aos dados cromatográficos. Os autores mostraram que o uso do R^2 é crucial para validar a precisão dos modelos empregados, garantindo resultados confiáveis e compatíveis com os realizados por humanos.

Conforme destacado por Gowd et al. (2018), a integração de lógica fuzzy com o coeficiente de determinação R^2 tem se mostrado promissora na identificação de padrões em dados. Essa pesquisa demonstrou que a combinação permite não só uma avaliação robusta, mas também a detecção de outliers e outros padrões em longas séries de dados. Recentemente, Ryoo et al. (2024) têm explorado os algoritmos de aprendizado de máquina na área da antidopagem, visando automatizar e otimizar as análises das amostras. Porém, diferente da nossa proposta, o trabalho utiliza amostras de sangue de atletas do gênero feminino em uma única modalidade esportiva. De qualquer modo, essas inovações afirmam reduzir o tempo de análise e aumentar a precisão dos resultados, representando um avanço significativo na área, no entanto, ainda carecem de estudos mais conclusivos.

Além disso, plataformas como o PubChem têm ganhado destaque como ferramentas de suporte em análises químicas. O PubChem, mantido pelo National Center for Biotechnology Information (NCBI), oferece um vasto repositório de informações químicas, incluindo estrutura molecular, propriedades físico-químicas e identificadores únicos como CAS. Segundo Kim et al. (2016), sua integração em análises cromatográficas permite enriquecer a interpretação dos dados, facilitando a caracterização de substâncias suspeitas e aumentando a confiabilidade dos resultados. Trabalhos como os de Zhang et al. (2021) demonstram como o uso de bases de dados químicas auxilia na identificação de padrões emergentes em estudos de antidopagem, indicando uma sinergia promissora com técnicas baseadas em lógica fuzzy e análise R^2.

A avaliação dos trabalhos relacionados indicou que a combinação de lógica fuzzy e coeficiente de determinação R^2, aliada ao uso de plataformas como o PubChem, apresenta elevado potencial na avaliação de cromatogramas de HPLC-MS. As pesquisas destacam não apenas a viabilidade dessa combinação, mas também apontam perspectivas para o desenvolvimento de técnicas mais sofisticadas para detecção de substâncias alvo

do controle de dopagem ao se incorporar conceitos de pipelines capazes de analisar grandes massas de dados de forma integrada e automatizada.

3. Materiais e métodos

O LBCD utiliza cromatógrafos líquidos de alta eficiência (HPLC) acoplados a equipamentos de espectrometria de massas de alta resolução (MS) da Thermo Scientific para realizar as corridas cromatográficas das amostras de urina fornecidas pelos atletas. Esses equipamentos são essenciais para a obtenção de dados científicos precisos sobre as substâncias presentes nas amostras a serem analisadas. Eles produzem os datasets utilizados nesta pesquisa, compostos por centenas de arquivos com dados textuais estruturados contendo as sequências de injeções das amostras (marchas analíticas validadas pelo LBCD e pelas normas ISO17.025 e ISL2021 da WADA), garantindo maior padronização, auditabilidade, reprodutibilidade e confiabilidade dos resultados. A sequência de injeção das amostras nos HPLC-MS para a geração dos datasets segue uma lógica que inclui injeções de amostras contendo as substâncias alvo (CQCL), injeções de solventes sem substâncias alvo (CQN), amostras de urina coletadas dos atletas e, ao final de cada corrida, a reinjeção do CQCL para comparação e verificação de ocorrências nas amostras.

As substâncias presentes nas amostras injetadas são separadas através do processo de cromatografia líquida, e sua detecção é feita por espectrometria de massas, gerando dados brutos armazenados em arquivos .RAW. Esses arquivos contêm informações detalhadas sobre parâmetros de análise, tempos de retenção (RT), concentrações das substâncias e outras variáveis cruciais para interpretação cromatográfica. Cada arquivo .RAW possui aproximadamente 15.800 registros e ocupa cerca de 100.00MB, representando um volume massivo de dados que precisa ser analisado. Os cromatogramas gerados são interpretados manualmente por analistas do LBCD, que comparam os dados obtidos com padrões de referência (CQCL). Este processo, embora rigoroso, é demorado, caro e suscetível a erros humanos devido ao grande volume de informações.

Para assegurar a veracidade das análises, cada cromatograma é interpretado de forma isolada por dois analistas. As interpretações podem resultar em uma conclusão de negativo, quando a substância alvo não está presente ou está em concentração muito baixa, ou presumível, quando há indícios da substância que necessitam de uma investigação mais detalhada. Em casos de divergência entre os analistas ou de conclusão presumível, novos métodos analíticos são aplicados para confirmação dos resultados.

Para abordar esses desafios, propomos uma metodologia baseada na teoria dos workflows científicos, implementados sob a forma de pipelines computacionais, para o processamento automatizado de grandes volumes de dados cromatográficos. Este pipeline contempla etapas que vão desde a injeção das amostras e geração dos arquivos .RAW até a análise automatizada dos dados utilizando lógica fuzzy e coeficiente de determinação R^2. A análise computacional compara os parâmetros extraídos dos cromatogramas das amostras com os padrões de CQCL, calculando o R^2 para indicar a presença ou ausência de substâncias alvo. A lógica fuzzy, por sua vez, interpreta qualitativamente os resultados obtidos, aproximando a análise computacional da realizada manualmente pelos analistas, mas com maior rapidez e menor suscetibilidade a erros.

Adicionalmente, o pipeline incorpora informações provenientes do PubChem, uma plataforma pública mantida pelo National Center for Biotechnology Information

(NCBI). O PubChem fornece dados detalhados sobre substâncias químicas, incluindo identificadores únicos (CAS), estrutura molecular, nomes sistemáticos e propriedades físico-químicas. A integração dessas informações no pipeline permite enriquecer a interpretação dos resultados cromatográficos ao associar os dados analíticos às características químicas das substâncias detectadas. Isso possibilita uma caracterização mais abrangente das substâncias presentes nas amostras e facilita a identificação de novos padrões de dopagem. Dessa forma, a combinação de análise automatizada dos cromatogramas com dados do PubChem potencializa a eficácia do processo de detecção e interpretação, fornecendo aos analistas uma ferramenta poderosa para reduzir o tempo de avaliação e aumentar a confiabilidade dos laudos emitidos.

4. Implementação

O pipeline, descrito na seção 3, foi materializado utilizando a linguagem de programação Python e suas bibliotecas, sendo capaz de realizar todas as computações relativas aos procedimentos analíticos. A primeira tarefa do pipeline é a carga dos datasets (dados do HPLC-MS com os arquivos .RAW). Em seguida, efetua a verificação e tratamento dos dados através da biblioteca pymsfilereader, desenvolvida por François [2019], onde a função GetChroData acessa os dados .RAW e extrai os parâmetros de corrida cromatográfica (StartTime, EndTime, MassRange, ScanFilter, SmoothingType e SmoothingValue), conforme descrito na Tabela 1.
Os parâmetros extraídos com a biblioteca pymsfilereader são cruciais para a análise dos dados que utilizam o tempo de retenção (RT) na coluna cromatográfica (definido pelos parâmetros StartTime e EndTime). A detecção da substância alvo é realizada utilizando os parâmetros MassRange e ScanFilter, que especificam o intervalo de massa e filtros de varredura para a análise cromatográfica. Esses dados estruturados em **DataFrames** (biblioteca pandas) representam os tempos de retenção e as intensidades de íons de cada massa molecular das substâncias em análise, organizando as gaussianas (picos dos cromatogramas) e facilitando a interpretação visual pelos analistas.
Para avaliar a similaridade entre os cromatogramas, utilizamos a função r2_score (da biblioteca scikit-learn) para calcular o R^2, que indica a qualidade do modelo de regressão. Essa análise é realizada comparando-se um conjunto de valores reais (y_true), extraídos dos cromatogramas de CQCL, com valores preditos (y_pred), extraídos das amostras de urina dos atletas. Em seguida, a lógica fuzzy, implementada com o módulo skfuzzy (da biblioteca scikit-fuzzy), é empregada para tratar incertezas nos dados, permitindo a classificação das amostras em categorias como: Negativo, Presumível Muito Baixo, Presumível Baixo, Presumível Médio, Presumível Alto ou Presumível Muito Alto.

Além das análises descritas, o pipeline utiliza a biblioteca pubchempy para a integração com a base de dados PubChem. A pubchempy é uma interface para acessar os dados químicos disponíveis na plataforma PubChem, permitindo consultas automatizadas para recuperar informações detalhadas sobre as substâncias identificadas nos cromatogramas. Com a biblioteca, é possível obter dados como nome IUPAC, identificador CAS, estrutura molecular, propriedades físico-químicas, solubilidade e toxicidade. Essas informações são utilizadas para enriquecer a análise, oferecendo subsídios adicionais para a interpretação dos resultados e facilitando a identificação de padrões emergentes de dopagem. A consulta é feita utilizando o identificador molecular (InChIKey) ou nomes comuns, retornando objetos estruturados que podem ser

manipulados diretamente no pipeline. Na tabela 1, a replicabilidade com as versões e suas bibliotecas.

Tabela 1. Bibliotecas e versões utilizadas no pipeline computacional

Biblioteca	**Versão**	**Descrição**	**Origem**
pubchempy	1.0.4	Interface para acesso à base de dados PubChem, utilizada para recuperação de dados químicos.	Externa
pymsfilereader	1.0.1	Ferramenta para leitura de arquivos .RAW provenientes de espectrometria de massas.	Externa
pandas	1.5.3	Biblioteca para manipulação e análise de dados estruturados em DataFrames.	Externa
matplotlib.pyplot	3.9.2	Ferramenta para criação de gráficos e visualização de resultados.	Externa
sklearn.metrics	1.5.2	Ferramenta para cálculo de métricas estatísticas, como o coeficiente de determinação R^2.	Externa
os	Padrão Python	Biblioteca padrão para operações com arquivos e diretórios no sistema.	Nativa do Python
numpy	1.24.0	Biblioteca para cálculo numérico eficiente e manipulação de arrays multidimensionais.	Externa

Biblioteca	Versão	Descrição	Origem
skfuzzy	0.5.0	Ferramenta para implementação de lógica fuzzy e sistemas de inferência.	Externa
timeit	Padrão Python	Biblioteca para medição do tempo de execução de trechos de código.	Nativa do Python
textwrap	Padrão Python	Biblioteca para manipulação e formatação de texto.	Nativa do Python
rdkit.Chem	2024.3.6	Ferramenta para manipulação de estruturas químicas e cálculos moleculares.	Externa
rdkit.Chem.Draw	2024.3.6	Módulo para criação de representações visuais de moléculas.	Externa
matplotlib.offsetbox	3.9.2	Submódulo do Matplotlib para manipulação avançada de gráficos, como inclusão de imagens.	Externa
mpl_toolkits.axes_grid1	3.9.2	Extensão do Matplotlib para criação de layouts complexos e inserção de gráficos adicionais.	Externa
pkg_resources	Padrão Python	Biblioteca para gerenciamento de dependências e pacotes instalados no ambiente Python.	Nativa do Python

Biblioteca	Versão	Descrição	Origem
Versão do Python	**3.10.0**	Ambiente de desenvolvimento utilizado para implementação do pipeline.	Nativa do Python

Os parâmetros são cruciais para a análise dos dados que utiliza o tempo de retenção (RT) na coluna cromatográfica através do pymsfilereader com *StartTime* e *EndTime*, a detecção da substância alvo será calculada através do *MassRange* e *ScanFilter*. Para, a seguir, organizar as gaussianas (picos dos cromatogramas) e agilizar a interpretação visual pelos analistas. Os parâmetros de *Smoothing* são responsáveis por ajustar os cromatogramas para visualização. Os dados são estruturados em *DataFrames* (biblioteca *pandas*) que representam os tempos de retenção e as intensidades de íons de cada massa molecular das substâncias em análise.

Para comparar os picos dos cromatogramas, utilizamos a função *r2_score* (da biblioteca *scikit-learn*) para calcular o R^2 que avalia a qualidade do modelo de regressão. Para determinar o R^2, utilizamos um conjunto de valores reais (*y_true*) extraídos dos arquivos de CQCL e um conjunto de valores preditos pelo modelo (*y_pred*), que são valores extraídos das amostras de urina dos atletas. A lógica *fuzzy* trata as incertezas e imprecisões presentes nos dados, ela é implementada através do módulo *skfuzzy* (da biblioteca *scikit-fuzzy*).

O *skfuzzy* oferece funcionalidades para criar sistemas de inferência *fuzzy*, que podem ser aplicados em diversas áreas, desde processos de tomada de decisão até o reconhecimento de padrões. O *pipeline* também utiliza funções de pertinência para o coeficiente, classificação, regras e variáveis de entrada e saída, com o objetivo de classificar cada amostra como sendo: *Negativo, Presumível Muito Baixo, Presumível Baixo, Presumível Médio, Presumível Alto* ou *Presumível Muito Alto*. Utilizamos a biblioteca *matplotlib.pyplot* para a visualização gráfica dos resultados, gerando gráficos complementares que auxiliam os analistas na interpretação dos cromatogramas e dos resultados das análises *fuzzy*.

Para complementar as análises químicas, a biblioteca **RDKit** é empregada no pipeline para a manipulação e visualização de estruturas moleculares. A **RDKit** permite a geração de representações visuais das moléculas em formatos como SMILES ou arquivos moleculares 2D/3D. Além disso, a biblioteca possibilita o cálculo de propriedades químicas relevantes, como massa molecular, polaridade, logP, e padrões de ligação. Essas funcionalidades são utilizadas para criar gráficos representativos das moléculas identificadas, auxiliando na caracterização das substâncias suspeitas de doping. A integração entre **pubchempy** e **RDKit** permite que o pipeline gere relatórios completos, incluindo não apenas os resultados analíticos, mas também a representação gráfica e os parâmetros químicos de cada substância, promovendo uma análise detalhada e de fácil interpretação.

Os experimentos envolveram a utilização integrada das bibliotecas pymsfilereader, pandas, matplotlib.pyplot, sklearn.metrics, numpy, skfuzzy, pubchempy e RDKit, proporcionando uma abordagem robusta e eficiente para analisar grandes quantitativos de amostras e compará-las com os controles positivos, culminando na classificação das amostras com base na lógica *fuzzy*. O código fonte dos *pipelines* está disponível em https://github.com/guyjunior/dopinho.

5. Resultados e discussões

Os experimentos computacionais buscaram 125 substâncias alvo presentes em 300 amostras cegas, 300 amostras de rotina e 20 amostras de CQCL. Foram realizadas várias rodadas de testes e os indicadores de acurácia, precisão e F1-score estão representados na Tabela 2. A 1ª rodada utiliza os dados com substâncias alvo, comparando a similaridade do CQCL com amostras fortificadas com todos as substâncias alvo. A 2ª rodada utiliza os dados cegos das amostras de rotina, comparando a similaridade do CQCL com amostras de rotina que os analistas interpretam. Por fim a 3ª rodada contém os dados com substâncias alvo e amostras cegas, comparando a similaridade do CQCL com amostras fortificadas e negativas, misturadas de modo proposital para verificação pelo *pipeline*.

Tabela 2. Indicadores das rodadas dos experimentos computacionais

	Acurácia	Precisão	F1-Score
1ª Rodada	98%	98%	99%
2ª Rodada	98%	98%	99%
3ª Rodada	86%	86%	93%

Os experimentos avaliaram mais de 600.000 dados das amostras (cada uma com 125 substâncias alvo com 8 parâmetros) relativos aos conjuntos de amostras de urina do controle de dopagem submetidas ao HPLC-MS.

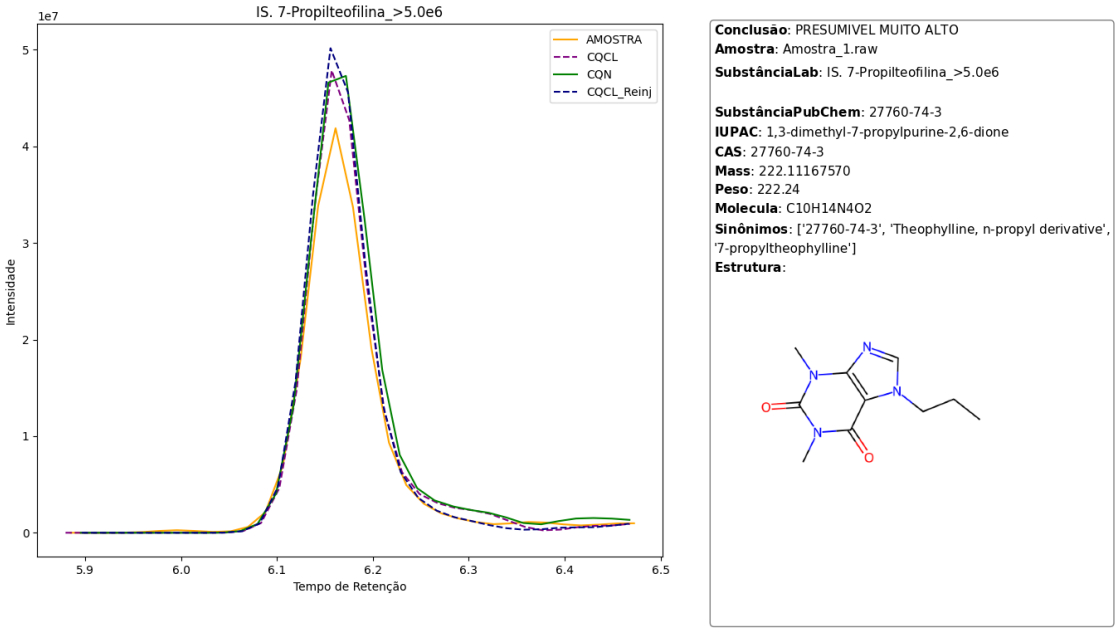

Figura 2. Alinhamento dos cromatogramas para validação com padrões internos e consulta ao PubChem.

A figura apresenta o alinhamento dos cromatogramas obtidos a partir das amostras em comparação com os padrões internos do laboratório. O alinhamento adequado é fundamental para garantir a confiabilidade dos resultados nas corridas cromatográficas subsequentes. A validação visual e estatística do alinhamento demonstra a consistência

dos tempos de retenção e a correta preparação das amostras, assegurando a padronização do processo analítico. Durante este processo, as substâncias detectadas no padrão interno são verificadas no PubChem para validação adicional, garantindo que suas propriedades químicas estejam em conformidade com os registros da base de dados.

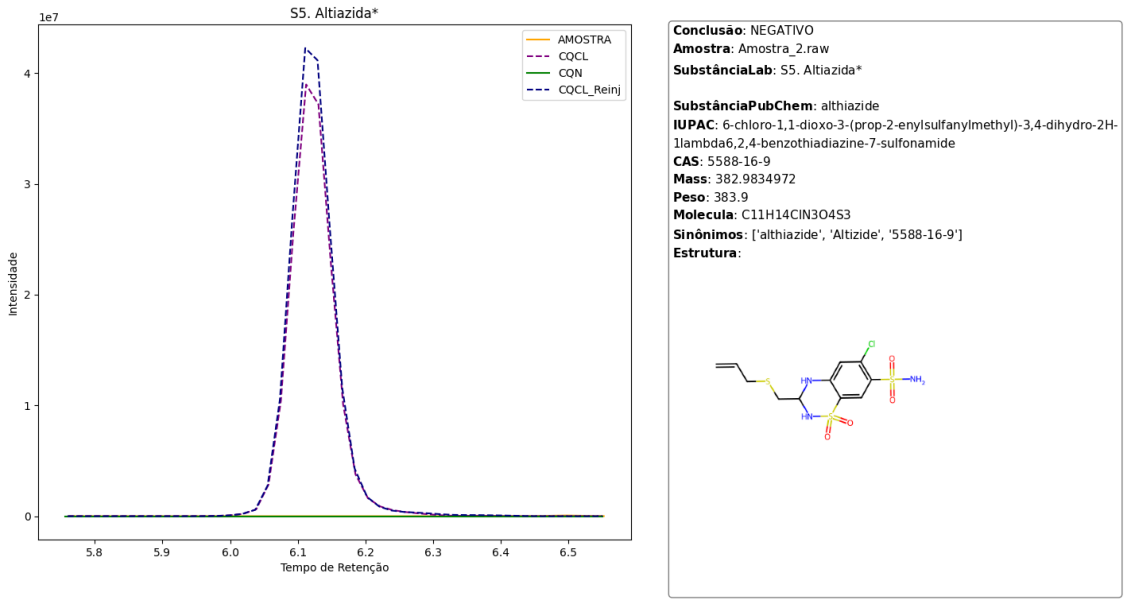

Figura 3. Classificação de resultados negativos e integração semântica com o PubChem.

A figura ilustra um exemplo de cromatogramas classificados como negativos, ou seja, amostras em que não foi evidenciada a presença de substâncias alvo. Por meio da quantificação do coeficiente R^2 e da lógica fuzzy, o resultado é interpretado de maneira não booleana, evidenciando maior flexibilidade e precisão na análise. Além disso, a figura demonstra a integração com o PubChem, exibindo informações químicas detalhadas sobre as substâncias analisadas, como estrutura molecular, propriedades físico-químicas e identificadores únicos (CAS). Essa integração enriquece o processo de interpretação e validação dos resultados.

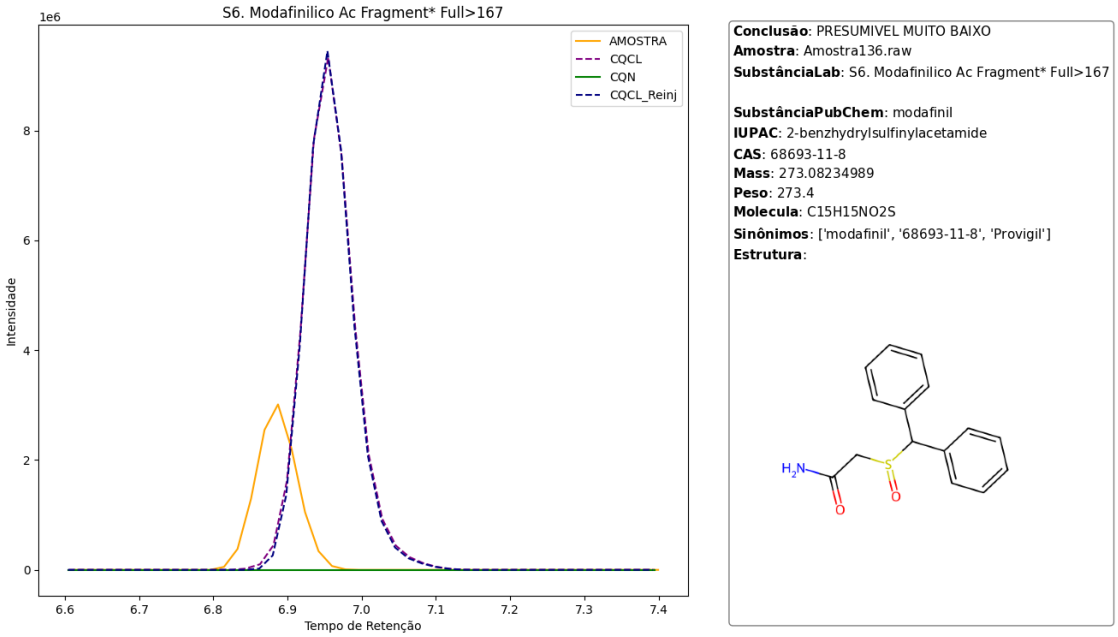

Figura 4. Classificação de resultados com presença presumível muito baixa e consulta ao PubChem.

Nesta figura, são apresentados cromatogramas em que a presença da substância alvo é classificada como presumível muito baixa. A lógica fuzzy utilizada no pipeline permite categorizar as amostras com base nos níveis de incerteza e concentração da substância detectada, oferecendo uma análise qualitativa robusta. Além disso, as informações químicas da substância, recuperadas do PubChem, são exibidas para complementar a análise, destacando propriedades como solubilidade e toxicidade, que podem ser úteis para interpretação.

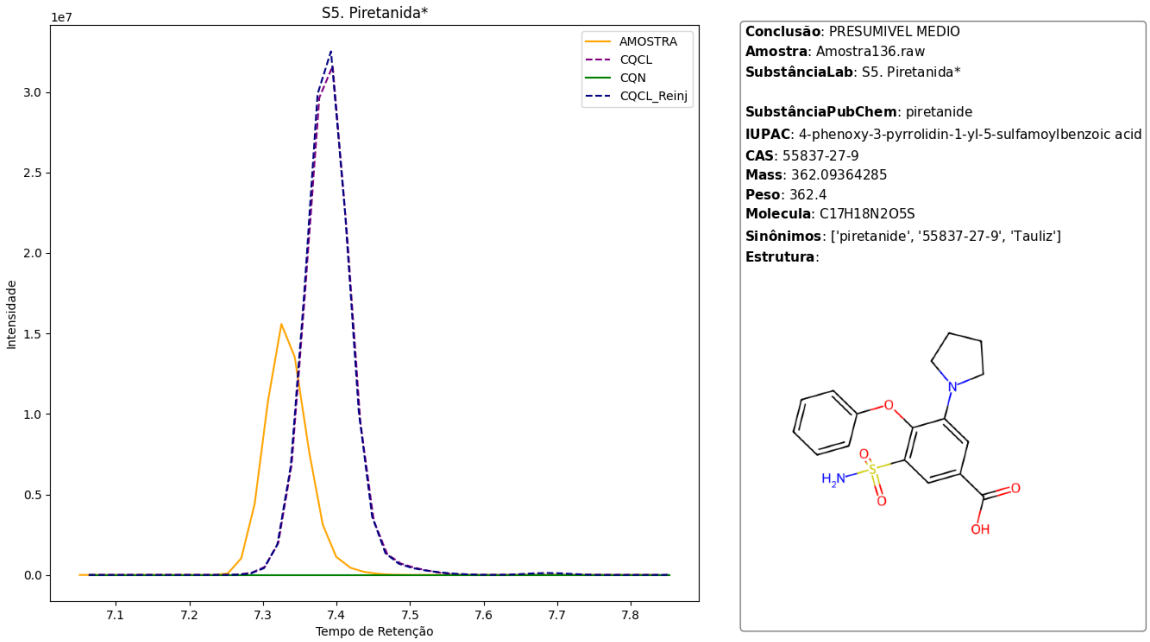

Figura 5. Classificação de resultados com presença presumível média e validação via PubChem

A figura exibe cromatogramas classificados como presumível médio. Esta classificação ocorre quando a concentração da substância alvo na amostra está em níveis intermediários. A análise combina a precisão quantitativa do R^2 com a lógica fuzzy, possibilitando uma categorização mais detalhada e confiável dos resultados. Adicionalmente, os dados químicos da substância são integrados ao resultado por meio do PubChem, permitindo uma análise contextual mais rica, com informações sobre aplicações e riscos associados.

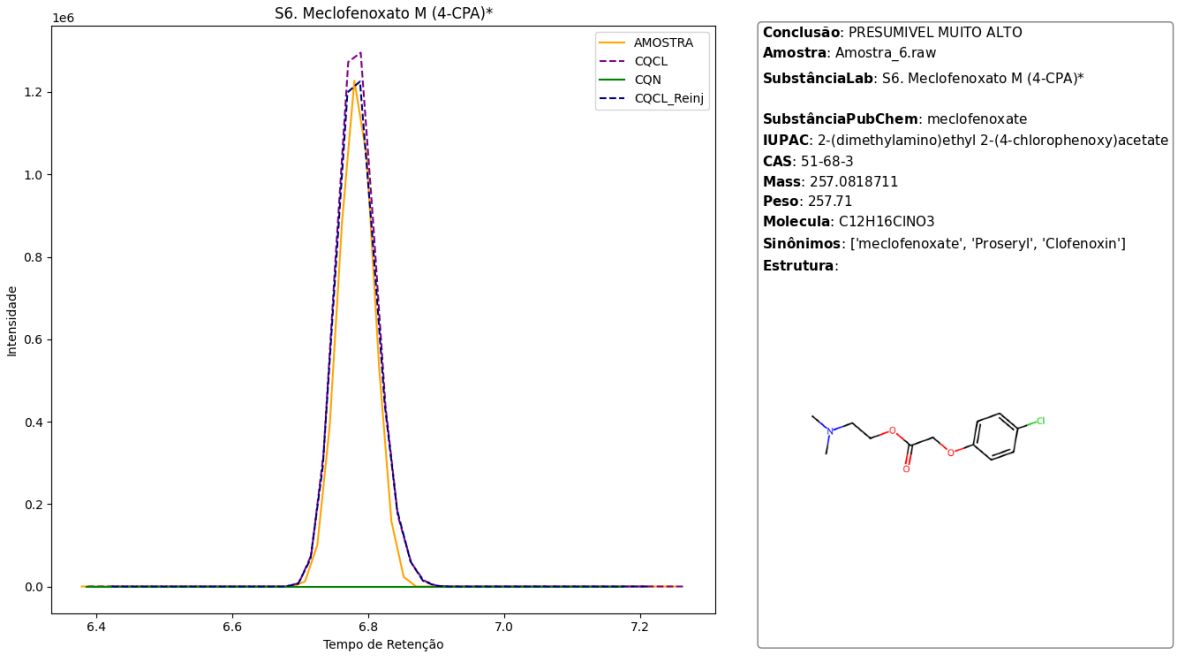

Figura 6. Classificação de resultados com presença presumível muito alta e integração detalhada ao PubChem

Esta figura apresenta cromatogramas onde a presença da substância alvo é classificada como presumível muito alta. A alta concentração da substância é claramente evidenciada pelos dados analíticos, sendo categorizada de forma precisa pela lógica fuzzy. Além disso, a integração com o PubChem fornece informações detalhadas sobre a substância detectada, incluindo estrutura molecular visualizada por meio do RDKit, propriedades físico-químicas e possíveis usos e restrições regulamentares. Essas informações auxiliam os analistas na emissão de laudos e na detecção de padrões emergentes de dopagem.

6. Conclusão

O controle de doping em atletas é um problema em aberto e de natureza interdisciplinar. Nossa contribuição, experimentos e seus resultados ainda são preliminares, no entanto, indicam que o método proposto tem indicativos de eficiência e precisão; atualmente, ele está sendo avaliado de forma experimental na rotina dos analistas do LBCD. A implementação computacional do método permitiu uma interpretação mais rápida dos dados cromatográficos quando comparada aos métodos tradicionais, evidenciando uma significativa redução no tempo de interpretação dos resultados sem perda de correção, o que sugere um aumento na confiabilidade dos resultados.

A capacidade do pipeline em processar grandes volumes de amostras em um curto período, aliada à precisão na identificação de substâncias alvo, está se mostrando plausível para o ambiente do controle de dopagem. A integração com o **PubChem**, permitindo a obtenção de informações detalhadas sobre as substâncias identificadas, como estrutura molecular, propriedades físico-químicas e classificações químicas, agrega uma camada adicional de suporte aos analistas. Essa funcionalidade enriquece a análise,

facilitando a caracterização das substâncias detectadas e possibilitando a identificação de novos padrões de dopagem.

A visualização facilitada de dados e a utilização da lógica fuzzy para a interpretação dos resultados apoiam os analistas no processo de tomada de decisões mais assertivas e seguras. A flexibilidade e a eficácia do método proposto indicam seu potencial para ser utilizado em outras áreas que requerem interpretação de dados complexos provenientes do HPLC-MS. Como trabalhos futuros, pretende-se ampliar o número de amostras, agregar novos métodos de aprendizado de máquina ao pipeline para classificação das amostras e oferecer interfaces gráficas mais intuitivas para os analistas. Além disso, busca-se expandir a integração com bases químicas externas, como o PubChem, para fornecer dados ainda mais completos. Também pretende-se oferecer a solução sob a licença de código livre para outros laboratórios de controle de dopagem ou setores que necessitem desse tipo de suporte computacional.

Referências

Abed, S. S., & Rasheed, A. S. (2020). Estimação simultânea de clonazepam e metronidazol em comprimidos farmacêuticos pelo modo de cromatografia líquida de alta eficiência de fase reversa com detecção uv. *Periódico Tchê Química*, *17*(36).

Aquino Neto, F. R. D. (2001). O papel do atleta na sociedade e o controle de dopagem no esporte. *Revista Brasileira de Medicina do Esporte*, *7*, 138-148.

Chiode, A. D. S. (2021). *Avaliação de propostas de coeficientes de determinação do tipo R^2 em modelos de regressão logística com resposta nominal* (Doctoral dissertation, Universidade de São Paulo).

François, A. (2019). *pymsfilereader: Thermo MSFileReader Python bindings*. GitHub. https://github.com/frallain/pymsfilereader

Gleaves, J., et al. (2021). Doping prevalence in competitive sport: evidence synthesis with "best practice" recommendations and reporting guidelines from the WADA Working Group on Doping Prevalence. *Sports Medicine*, *51*(9), 1909-1934.

Gowd, B. P., Jayasree, K., & Hegde, M. N. (2018). Comparison of artificial neural networks and fuzzy logic approaches for crack detection in a beam like structure. *Int. J. Artif. Intell. Appl*, *9*(1), 35-51.

Maennig, W. (2014). Inefficiency of the anti-doping system: Cost reduction proposals. *Substance use & misuse*, *49*(9), 1201-1205.

Mattoso et al. (2010). Towards supporting the life cycle of large scale scientific experiments. *Int. J. of Business Process Intehration and Management*. 5(1), 79-92.

Marro, A. A., et al. (2010). Lógica fuzzy: conceitos e aplicações. *Natal: Universidade Federal do Rio Grande do Norte (UFRN)*, 2.

Mogollon, N. G., et al. (2014). State of the art two-dimensional liquid chromatography: fundamental concepts, instrumentation, and applications. *Química Nova*, *37*, 1680-1691.

Pereira, S. L. R. (2022). *Desenvolvimento e validação de um método de detecção e quantificação de THC-COOH, em urina, por cromatografia gasosa acoplada a espectrometria de massa* (Doctoral dissertation).

Ryoo, H., et al. (2024). Identification of doping suspicions through artificial intelligence-powered analysis on athlete's performance passport in female weightlifting. *Frontiers in Physiology*, *15*, 1344340

Ukić, Š., et al. (2015). Development of gradient retention model in ion chromatography. Part III: Fuzzy logic QSRR approach. *Chromatographia*, *78*, 889-898.Dias, M. D. S., Visintin, L., & Reiser, R. Estudo Introdutório da Lógica Fuzzy Intuicionista Intervalar.