# ChromLake: Plataforma para Interoperabilidade e Reuso de Dados Cromatográficos

Luiz Eduardo D. Paiva[1], Guy M. B. Junior[1], Giseli Rabello Lopes[1], Jorge Zavaleta[1],[2], Sergio Manuel Serra da Cruz[1]

[1] Programa de Pós-Graduação em Informática (UFRJ) – Rio de Janeiro – RJ – Brasil

[2] Alda Research Institute (ALDA) – Rio de Janeiro – RJ – Brasil

ledp@iq.ufrj.br, giseli@ic.ufrj.br, sergioserra@ic.ufrj.br, zavaleta.jorge@gmail.com, guyjunior@iq.ufrj.br

***Abstract.*** *ChromLake is an innovative computational strategy that enables the access, transformation, and storage of large volumes of data generated by various types of analytical equipment. ChromLake mitigates recurrent limitations in high-demand laboratory environments and supports the automation of distinct sample analyses with a high degree of reliability. The proposal is based on pipelines that provide data interoperability, persistence, and security. Experimental results indicate that the strategy preserves chromatographic data without detectable losses in the experiments performed.*

***Resumo.*** *ChromLake é uma plataforma computacional inovadora que provê acesso, transformação e armazenamento de grandes volumes de dados produzidos por diversos tipos de equipamentos analíticos. ChromLake mitiga as limitações recorrentes em ambientes laboratoriais de alta demanda e apoia a automação das análises de amostras distintas com alto grau de confiabilidade. A estratégia baseia-se em pipelines que oferecem interoperabilidade, persistência e rastreabilidade de dados. Os primeiros resultados experimentais indicam que a estratégia preserva os dados cromatográficos sem perdas detectáveis nos experimentos realizados.*

## 1. Introdução

A Química Analítica contemporânea tem sido impulsionada pelo avanço de técnicas instrumentais de elevada sensibilidade e seletividade. Nesse cenário, o emprego de equipamentos de Cromatografia Líquida de Alta Eficiência acoplada à Espectrometria de Massas de Alta Resolução (LC-HRMS) e da Cromatografia Gasosa acoplada à Espectrometria de Massas (GC-MS) tornou-se indispensável para a caracterização de amostras complexas [Heinsvig et al. 2023; Schymanski et al. 2014]. Esses equipamentos realizam corridas cromatográficas (*scans*) produzindo cromatogramas baseados em grandes datasets que expressam propriedades físicas das substâncias, como tempos de retenção (RT), que indicam o instante em que cada substância elui da coluna cromatográfica, razões massa/carga (m/z), que identificam os íons detectados no espectrômetro, e intensidades de sinais [Mardal et al. 2023].

Os *scans* são fundamentais para identificação de padrões, detecção de metabólitos e investigações retrospectivas de doping e contaminações [Mogollón et al. 2014]. Entretanto, os dados cromatográficos são produzidos em formatos proprietários dependentes do fornecedor dos equipamentos, limitando a interoperabilidade, restringindo o acesso programático e onerando a construção de workflows orientados a análises preditivas [Mattoso et al. 2010; Allain 2019]. Na prática, as análises dependem frequentemente de acesso físico aos equipamentos e de procedimentos manuais exaustivos, tornando o processo pouco escalável e inadequado para ambientes de alta demanda como

os laboratórios de controle de doping homologados pela Agência Mundial Antidopagem (WADA) [Magalhães Britto Junior et al. 2024].

Segundo Mogollón et al. (2014) e Milman e Zhurkovich (2017), há grande potencial para o desenvolvimento de soluções computacionais inovadoras voltadas para a área da cromatografia, mas sua disseminação ainda é limitada na academia e na indústria. Além disso, as ferramentas existentes concentram-se majoritariamente na análise de scans isolados, sem tratar de forma sistemática o ciclo de vida dos dados em repositórios estruturados, alinhados aos princípios de dados FAIR [Wilkinson et al. 2016].

Este artigo apresenta a plataforma ChromLake. Ela foi concebida para facilitar o acesso, estruturação, armazenamento e análises de grandes quantidades de arquivos brutos de scans cujos dados são persistidos em banco de dados colunares de processamento analítico online (OLAP) e acessíveis via consultas em linguagem SQL. A plataforma adota técnicas de e-science e engenharia de dados, assegurando persistência, escalabilidade e rastreabilidade em um repositório estruturado, rastreável e independente de fabricante e do tipo de amostra analisada.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, a Seção 3 descreve os materiais e métodos, a Seção 4 apresenta a implementação, a Seção 5 discute os resultados e a Seção 6 reúne as considerações finais.

## 2. Trabalhos Relacionados

Mardal et al. (2023) indicam a viabilidade do uso de SGBDs relacionais para o arquivamento e análise escalável de dados de LC-HRMS por meio do ScreenDB, eficaz em análises retrospectivas de mais de 40.000 arquivos. Entretanto, sua implementação foca em modelos relacionais e fluxos específicos para triagens não-direcionadas. Embora o ChromLake também serialize os formatos proprietários para mzML por meio de extratores especializados, seu principal diferencial não reside na independência de fornecedor, mas no ganho de desempenho e escalabilidade proporcionado por um SGBD OLAP colunar, com indexação física, projeções de pré-agregação e compressão nativa orientadas a consultas analíticas. Outra abordagem, o Metabolite Atlas de Yao et al. (2015), integra dados de LC/MS em clusters HPC com SciDB, realizando operações de *data slicing* sobre matrizes multidimensionais (RT × m/z × Intensidade). No entanto, seu foco está na exploração interativa de metabólitos, requer linguagem AQL/AFL, não provê independência de fabricante nem suporta GC-MS.

Do ponto de vista arquitetural, o ChromLake alinha-se aos conceitos de *Data Lake* (DL) discutidos por Miloslavskaya e Tolstoy (2016) e Milman e Zhurkovich (2017), que defendem repositórios seguros e escaláveis capazes de manter dados em formato nativo para transformações orientadas à análise (*schema-on-read*). Ao serializar arquivos proprietários para mzML e persisti-los em ambiente colunar, o ChromLake implementa uma estratégia de repositório estruturado que combina fidelidade e rastreabilidade de um DL com o desempenho necessário para grandes volumes de dados espectrométricos.

Bi et al. (2026) desenvolveram o DCOP, uma plataforma de acesso aberto com mais de 214 mil compostos organofosforados e predição de espectros MS/MS via redes neurais. Embora relevante para curadoria química e predição *in silico*, o DCOP não aborda a ingestão de dados cromatográficos brutos em repositórios analíticos. No campo forense, Pasternak et al. (2022) demonstraram com o OpenChrom que sistemas automatizados podem atingir 100% de concordância com especialistas na detecção de líquidos inflamáveis em 374 amostras de GC-MS. Contudo, o trabalho é centrado no processamento de arquivos individuais. O ChromLake complementa esse ecossistema ao focar na camada de infraestrutura e engenharia de dados, viabilizando ingestão automatizada de múltiplos fabricantes e tipos de amostras em repositórios OLAP com persistência estruturada em SQL.

Em síntese, os trabalhos relacionados avançam no uso de bancos de dados e formatos abertos, mas tendem a contextos específicos, restritos a uma técnica ou fabricante, ou apoiados em modelos de armazenamento menos otimizados para consultas analíticas em larga escala. O ChromLake diferencia-se ao combinar um método geral de engenharia de dados, aplicável a múltiplas técnicas e fabricantes, com uma camada de persistência colunar OLAP projetada para alto desempenho e escalabilidade, priorizando reprodutibilidade e reuso dos dados.

## 3. Material e Métodos

Os dados brutos utilizados nessa pesquisa são gerados por equipamentos previamente instalados, homologados e calibrados do Laboratório Brasileiro de Controle de Dopagem (LBCD/UFRJ). Os datasets de LC-HRMS dizem respeito a scans de amostras de urina do (LBCD/UFRJ). Os datasets de GC-MS dizem respeito a scans de amostras de petróleo obtidas a partir de projetos conduzidos no Núcleo de Análises Forenses (NAF/UFRJ). Devido a questões legais e contratuais, os datasets foram utilizados unicamente para fins de validação metodológica e de funcionalidades, sem identificação dos fornecedores das amostras e indivíduos.

ChromLake é composta por macrotarefas: (1) validação, carga, verificação e serialização dos arquivos brutos para o formato mzML, (2) extração de dados cromatográficos e metadados sobre as corridas cromatográficas e configuração dos equipamentos, (3) ingestão de dados e metadados em repositório colunar e (4) reconstrução e visualização de cromatogramas através de recuperação via consultas SQL.

Na primeira macrotarefa os arquivos brutos produzidos pelo LC-HRMS com extensão .RAW (formato proprietário da empresa Thermo) ou produzidos pelo GC-MS com extensão .D (formato proprietário da empresa Agilent) são serializados para documentos XML (formato mzML) preservando todas as informações e metadados sem filtragens ou alteração dos dados brutos. Na segunda macrotarefa, são extraídas informações relativas aos RT, m/z, intensidades dos sinais e demais metadados de cada *scan*. A terceira macrotarefa ingere essas informações em um banco de dados colunar, para o qual se adotou o ClickHouse.

Na quarta macrotarefa, os dados são recuperados do banco através de consultas SQL e utilizados na reconstrução dinâmica e visualizações dos dados através de diversos tipos de cromatogramas, por exemplo, TIC (*Total Ion Chromatogram*), XIC (*Extracted Ion Chromatogram*) e EM (espectros de massa). Os TIC representam a soma da intensidade de todos os íons das substâncias detectadas (analitos) pelos equipamentos, os XIC limitam-se a representar a faixa de massa usada para monitoramento de analitos específicos.

O repositório ChromLakeDB foi implantado em uma instância do ClickHouse 25.7 executada como contêiner Docker sobre WSL2 (16 threads e cerca de 31 GiB de memória), em um computador com processador AMD Ryzen 7 5700G (8 núcleos, 16 threads), 64 GB de RAM DDR4, armazenamento SSD NVMe e Windows 11 Pro. Todos os experimentos de (desempenho e validação) foram realizados nessa configuração.

## 4. Plataforma ChromLake

A plataforma é composta por *pipelines* codificados em Python 3.10, um DL e um repositório estruturado de dados colunares, organizados na Figura 2 em quatro grupos: Dados Gerados, Processamento, Banco de Dados e Visualização.

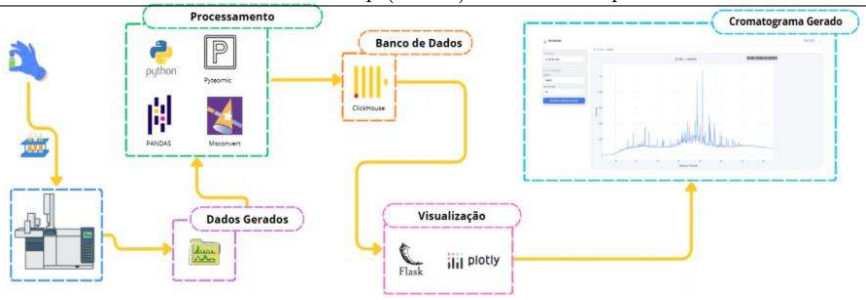


**Figura 2. Representação simplificada da implementação do ChromLake.**

O primeiro *pipeline*, que compõe o grupo Processamento da Figura 2, estabelece a conexão com os equipamentos para transferir os dados brutos (grupo Dados Gerados) para o DL local, em seguida, aciona atividades que detectam automaticamente o tipo de arquivo a ser convertido com base na sua extensão (.RAW ou .D). A seguir, inicia e controla programaticamente o fluxo de serializações através do *msconvert* do *toolkit* ProteoWizard e do *parser* que utiliza as bibliotecas *Pyteomics*, *Pandas* e *NumPy* para tratar dados brutos (textuais e numéricos) e estruturá-los em arquivos temporários em formato mzML no DL. A conversão para mzML é realizada pelo *msconvert*, ferramenta do *toolkit* ProteoWizard amplamente adotada como padrão da comunidade para a interoperabilidade de dados de espectrometria de massas [Chambers et al. 2012], e preserva todas as informações dos arquivos originais, sem filtragem, suavização ou centroidização.

A validação da fidelidade da engenharia de dados é realizada comparando-se, ponto a ponto, os dados presentes no arquivo mzML de origem com os dados equivalentes recuperados do ChromLakeDB via consultas SQL. Como a conversão dos arquivos brutos para mzML é assegurada pelo *msconvert*, ferramenta consolidada e validada pela comunidade [Chambers et al. 2012], o mzML é adotado como referência, restando comprovar que as etapas de extração, ingestão e recuperação preservam integralmente os dados. Para cada amostra são calculados dois indicadores complementares: o hash criptográfico SHA-256 dos vetores, que comprova identidade bit-a-bit, e o erro máximo absoluto, que quantifica a maior divergência ponto a ponto. A comparação é implementada com as bibliotecas *SciPy*, *NumPy* e *hashlib*.

O segundo *pipeline*, também pertencente ao grupo Processamento, recupera os arquivos mzML do DL e os processa para extrair os principais atributos analíticos para cada substância analisada. A extração recupera os dados de RT, razões m/z, intensidades e metadados de cada *scan* e equipamento.

O terceiro *pipeline*, correspondente ao grupo Banco de Dados, consiste na ingestão de dados e metadados de cada arquivo mzML no *schema* ChromLakeDB, implantado no ClickHouse, um banco de dados colunar OLAP executado em contêiner Docker. O *ChromLakeDB* é composto por cinco tabelas que cobrem dois cenários de aquisição. As tabelas de dimensão *samples* (metadados da amostra: identificador, nome, lote, técnica, equipamento e ano) e *sample_channels* (canais de aquisição, que definem o *channel_id* das séries temporais) utilizam o *engine* ReplacingMergeTree, que garante ingestão idempotente. As tabelas de série temporal utilizam o *engine* MergeTree, particionado por ano.

O modelo ramifica-se em dois caminhos a partir de *sample_channels*, conforme o tipo de corrida. Para corridas em modo *full scan* (LC-HRMS), o caminho é *samples* → *sample_channels* → *scans* (cabeçalho das varreduras, com tempo de retenção, sem m/z) → *peaks* (pares massa/carga e intensidade). Para corridas sem espectro associado (GC-MS em modo SIM, fragmentação de íon ou detecção FID), o caminho é *samples* → *sample_channels* → *chromatogram_points* (tempo de retenção e intensidade). Todos os relacionamentos são do tipo um-para-muitos (1:N), estabelecidos logicamente por colunas

compartilhadas, já que o ClickHouse não impõe chaves estrangeiras formais. A Figura 3 apresenta o diagrama do esquema.

Para acelerar a reconstrução de cromatogramas de íons extraídos (XIC) sobre grandes volumes de amostras, a tabela *peaks* emprega duas otimizações. A coluna inteira *mz_bin* converte a busca por massa em um filtro sobre inteiros, mais eficiente que comparações em ponto flutuante. A projeção *eic_by_rt* pré-agrega a soma das intensidades por canal, faixa de massa e tempo de retenção, permitindo resolver a consulta diretamente a partir dos dados pré-somados, sem varrer todos os picos nem executar junções. Adicionalmente, codecs especializados de compressão (*Gorilla* para intensidades, *T64* para inteiros, *ZSTD* e *LowCardinality*) reduzem o volume em disco e o custo de leitura, sustentando o desempenho das consultas analíticas.

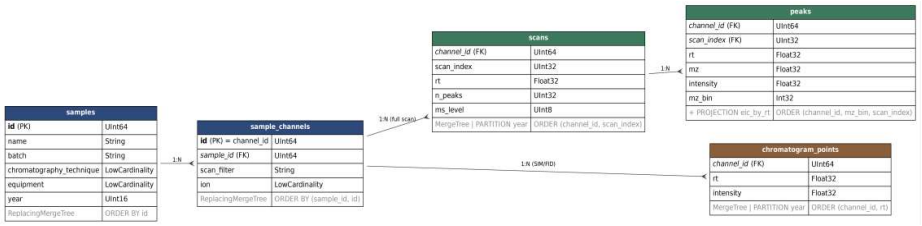


**Figura 3. *Schema* do ChromLakeDB.**

A ingestão dos dados no ChromLakeDB é mediada pela biblioteca *clickhouse-connect* que controla a ingestão e oferece recurso de repetição automática em caso de falhas. Ressaltamos que os fluxos de ingestão variam conforme a técnica analítica ou equipamento utilizado. Por exemplo, para equipamentos GC-MS operando em modo de monitoramento seletivo de íons (*Selected Ion Monitoring*, SIM) são processados cromatogramas específicos por íons. Já para os equipamentos LC-HRMS, são ingeridos cromatogramas e varreduras espectrais, incluindo registros MS e MS2. Para manter a correspondência e precisão numérica dos arquivos mzML, todos os dados armazenados no ChromLakeDB utilizam os mesmos tipos de dados dos arquivos de origem.

Por fim, o quarto *pipeline*, correspondente ao grupo Visualização, é acionado a partir de uma interface Web (Flask) e executa as consultas SQL submetidas pelos pesquisadores, retornando os dados solicitados para reconstrução e visualização dos diferentes tipos de cromatogramas e *scans*. A interface do ChromLake oferece a visualização de cromatogramas e espectros reconstituídos através da biblioteca *Plotly*.

## 5. Resultados Preliminares e Discussão

Os experimentos avaliaram 49 amostras distintas, sendo 29 de urina analisadas por LC-HRMS (Thermo, em modo *full scan*) e 20 de petróleo analisadas por GC-MS (Agilent, sendo cinco em modo *full scan* e quinze em modo SIM), totalizando mais de 63 milhões de pontos de dados. O requisito essencial para qualquer estratégia de análise em Química Analítica é a reprodutibilidade dos experimentos e a confiabilidade dos dados. Como estratégia de validação das macrotarefas, realizou-se a ingestão de todas as amostras e a comparação ponto a ponto entre os dados presentes nos arquivos mzML de origem e os dados recuperados do ChromLake, nas dimensões de intensidade, tempo de retenção (RT) e, quando aplicável, razão massa/carga (m/z). Foram empregados dois indicadores: o hash SHA-256, que comprova identidade bit-a-bit, e o erro máximo absoluto, que quantifica a maior divergência observada. Em nenhuma das 49 amostras houve divergência no número de pontos, descartando qualquer perda de dados na ingestão.

Nas 29 amostras de urina (LC-HRMS), os valores de intensidade e de razão massa/carga apresentaram erro máximo absoluto igual a zero, com hashes idênticos, confirmando preservação bit-a-bit. O tempo de retenção, armazenado em Float32 por eficiência de espaço, apresentou erro máximo absoluto da ordem de $4{,}8 \times 10^{-7}$, equivalente

à identidade dentro da precisão do formato. Em Float32, os hashes coincidiram em todas as amostras. Nas 20 amostras de petróleo (GC-MS, em modo full scan e SIM), os valores de intensidade também apresentaram erro máximo absoluto igual a zero, com hashes idênticos. O tempo de retenção e a razão massa/carga, em Float32, apresentaram erros máximos da ordem de $3{,}8 \times 10^{-6}$ e $2{,}4 \times 10^{-5}$, respectivamente, igualmente compatíveis com a precisão do formato. Em todas as 49 amostras avaliadas, os hashes em Float32 foram idênticos aos do mzML de origem.

**Tabela 1. Validação de fidelidade entre o arquivo mzML de origem e os dados recuperados do ChromLakeDB.**

| Técnica (modo) | Amostras | Hash SHA-256 idêntico | Erro máx. intensidade | Erro máx. RT | Erro máx. m/z |
|---|---|---|---|---|---|
| LC-HRMS (full scan) | 29 | 100% (29/29) | 0 | $4{,}8 \times 10^{-7}$ | 0 |
| GC-MS (full scan) | 5 | 100% (5/5) | 0 | $3{,}8 \times 10^{-6}$ | $2{,}4 \times 10^{-5}$ |
| GC-MS (SIM) | 15 | 100% (15/15) | 0 | $3{,}8 \times 10^{-6}$ | N/A |

A intensidade é preservada com erro absoluto igual a zero (identidade bit-a-bit). RT e m/z são armazenados em Float32, com resíduo no nível da precisão do formato. O hash SHA-256 em Float32 é idêntico ao do mzML de origem em todas as 49 amostras.

Os resultados indicam que o ChromLake preserva a fidelidade dos dados, sem perdas detectáveis introduzidas pelos pipelines nos experimentos realizados.

Para avaliar os ganhos de desempenho da plataforma, o ChromLake foi comparado ao fluxo tradicional dos softwares proprietários (Xcalibur para LC-HRMS e MassHunter para GC-MS), cujos tempos de operação foram informados por analistas experientes dos laboratórios. A avaliação considerou uma carga de mil amostras de urina (LC-HRMS) e cargas equivalentes de petróleo (GC-MS em modo full scan e SIM), abrangendo o tempo de reconstrução de cromatogramas, o custo de ingestão, a capacidade de consultas multi-amostra, a eficiência de armazenamento, a acessibilidade e a compressão. A Tabela 2 consolida os resultados.

**Tabela 2. Comparações entre os fluxos tradicional e o ChromLake.**

| Característica / Métrica | Fluxo Tradicional (Xcalibur) | Fluxo Tradicional (MassHunter) | ChromLake |
|---|---|---|---|
| Reconstrução de TIC por amostra [1] | 2,5 min | 1 min | LC-HRMS: 0,042 s; GC-MS: 0,010 a 0,135 s |
| Aceleração (speedup) [1] | N/A | N/A | 445× a 5.882× conforme a técnica |
| Ingestão por amostra [2] (custo único) | N/A | N/A | 0,09 a 15,7 s conforme a técnica |
| Consultas multi-amostra | Inviável | Inviável | 17.990 XICs sobre mil amostras de urina em 4,5 s |
| Armazenamento (LC-HRMS) | 92 MB/amostra | N/A | 13 MB/amostra (redução de 86%) |
| Armazenamento (GC-MS, full scan) | N/A | 38 MB/amostra | 27,6 MB/amostra (redução de 27%) |
| Armazenamento (GC-MS, SIM) | N/A | 3 MB/amostra | 0,2 MB/amostra (redução de 93%) |
| Acessibilidade dos dados | Dependente de software proprietário | Dependente de software proprietário | Consultável via SQL sem descompressão |
| Compressão de dados | Formato nativo/proprietário | Formato nativo/proprietário | Codecs especializados (Gorilla, T64, ZSTD, LowCardinality) |

[1] A reconstrução de TIC por amostra corresponde à mesma tarefa executada pelo analista humano nos softwares proprietários (abrir e reconstruir o cromatograma de uma amostra). No ChromLake, essa operação é replicável graças aos dados já ingeridos. Os ganhos por técnica são de 3.602× (LC-HRMS), 5.882× (GC-MS em modo SIM) e 445× (GC-MS em modo full scan).

[2] A ingestão é um custo único por amostra; todas as consultas subsequentes são realizadas sobre os dados persistidos.

Em síntese, os resultados experimentais indicam que o ChromLake reconstrói os cromatogramas preservando a fidelidade em relação aos dados de origem, com acelerações de centenas a milhares de vezes sobre o fluxo tradicional (de 445× a 5.882× conforme a técnica) e redução de armazenamento entre 27% e 93%. A ingestão, custo pago uma única vez, viabiliza consultas multi-amostra impraticáveis nas ferramentas proprietárias, como a extração de milhares de XICs sobre mil amostras em poucos segundos, organizando dados analíticos históricos em um repositório estruturado, independente de fabricante e com suporte a consultas SQL escaláveis e reprodutíveis.

## 6. Considerações Finais

Este trabalho apresentou o ChromLake, plataforma aberta para acesso, estruturação e armazenamento de dados cromatográficos, que pode ser utilizada em amostras de naturezas distintas. A validação dos scans evidenciou a preservação dos dados por meio de hashes idênticos e erro máximo absoluto compatível com a precisão do formato de armazenamento, indicando que os pipelines preservam os dados brutos sem perdas detectáveis nos experimentos realizados.

ChromLake endereça uma lacuna, ele oferece uma infraestrutura independente de fornecedor baseada nas melhores práticas da engenharia de dados, permitindo consultas rápidas sobre dados históricos e suportando triagens retrospectivas, controle de qualidade em nível de lote e apoio ao desenvolvimento de novas aplicações de IA em larga escala.

Como perspectivas futuras, destacam-se a expansão a outros fabricantes e técnicas analíticas, aprimoramento das interfaces Web permitindo insights mais profundos e implementação de novos pipelines para coletar metadados mais ricos e proveniência retrospectiva (padrão PROV da W3C) sobre as execuções dos experimentos e persistência no repositório em conformidade com as normas internacionais (e.g. GLP, GMP ou 21 CFR Part 11), permitindo que os laboratórios aproveitem o potencial de seus dados para orientar decisões e impulsionar a IA.

## Agradecimentos

Este trabalho foi realizado com apoio da CAPES - Código de Financiamento [001]; também contou com apoios parciais da ALDA e do CNPq - Processos nº [310452/2025-2] e da FAPERJ - Processo nº [E-26/260.253/2026]. Os autores agradecem também à Agência Mundial Antidopagem (WADA) pelo financiamento do projeto DopingLake (Projeto nº [261F01HP]), aprovado no ciclo 1 de 2026, que aplica a metodologia apresentada neste trabalho. Agradecem ainda ao Instituto Nacional da Propriedade Industrial (INPI) pelo registro de software do ChromLake (Processo nº [BR512026002858-2]).

## Referências

Allain, F. pymsfilereader: Thermo MSFileReader Python bindings. [S. l.]: GitHub, 2019. Disponível em: https://github.com/frallain/pymsfilereader. Acesso em: 4 abr. 2026.

Bi, R. et al. Database of Chemicals for Organophosphorus (DCOP): An Information Application Platform for Organophosphorus Compounds. *Analytical Chemistry*, 2026. DOI: 10.1021/acs.analchem.5c06885.

Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, v. 30, n. 10, p. 918-920, 2012. DOI: 10.1038/nbt.2377.

Heinsvig, P. J. et al. Forensic drug screening by liquid chromatography hyphenated with high-resolution mass spectrometry (LC-HRMS). *TrAC Trends in Analytical Chemistry*, v. 162, p. 117023, 2023. DOI: 10.1016/j.trac.2023.117023.

Magalhães Britto Junior, G. et al. Aplicando Lógica Fuzzy na Interpretação de Grandes Volumes de Dados Cromatográficos no Controle de Dopagem. In: BRAZILIAN E-

SCIENCE WORKSHOP (BRESCI), 18., 2024, Florianópolis/SC. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 24-31. DOI: 10.5753/bresci.2024.244261.

Mardal, M. et al. Scalable Analysis of Untargeted LC-HRMS Data by Means of SQL Database Archiving. *Analytical Chemistry*, v. 95, n. 10, p. 4592-4596, 2023. DOI: 10.1021/acs.analchem.2c03769.

Mattoso, M. et al. Towards supporting the life cycle of large scale scientific experiments. *International Journal of Business Process Integration and Management*, v. 5, n. 1, p. 79-92, 2010. DOI: 10.1504/IJBPIM.2010.033176.

Milman, B. L.; Zhurkovich, I. K. The chemical space for non-target analysis. *TrAC Trends in Analytical Chemistry*, v. 97, p. 179-187, 2017. DOI: 10.1016/j.trac.2017.09.013.

Miloslavskaya, N.; Tolstoy, A. Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*, v. 88, p. 300-305, 2016. DOI: 10.1016/j.procs.2016.07.439.

Mogollón, N. G. S. et al. O estado da arte da cromatografia líquida bidimensional: conceitos fundamentais, instrumentação e aplicações. *Química Nova*, v. 37, n. 10, p. 1680-1691, 2014. DOI: 10.5935/0100-4042.20140261.

Pasternak, Z. et al. Automatic detection and classification of ignitable liquids from GC-MS data of casework samples in forensic fire-debris analysis. *Forensic Chemistry*, v. 29, p. 100419, 2022. DOI: 10.1016/j.forc.2022.100419.

Schymanski, E. L. et al. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environmental Science & Technology*, v. 48, n. 3, p. 1811-1818, 2014. DOI: 10.1021/es4044374.

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, v. 3, p. 160018, 2016. DOI: 10.1038/sdata.2016.18.

Yao, Y. et al. Analysis of Metabolomics Datasets with High-Performance Computing and Metabolite Atlases. *Metabolites*, v. 5, n. 3, p. 431-442, 2015. DOI: 10.3390/metabo5030431.