

# Projeto Massa

Douglas Galvão Machado<sup>1</sup>, Gabriel Ramos Ferreira<sup>1</sup>, João Pedro Silva Braga<sup>1</sup>,  
João Vitor Romero Sales<sup>1</sup>, Lucas Randazzo<sup>1</sup>

<sup>1</sup>Instituto de Ciências Exatas e Informática  
Pontifícia Universidade de Minas Gerais (PUC Minas)  
Belo Horizonte – MG – Brasil

{dgsmachado, gabriel.ferreira.732131, joao.braga.1463240,  
joao.sales.7851111, lucas.randazzo }@sga.pucminas.br

**Resumo.** O algoritmo MASSA ("Molecular dAta Set SAmpling Algorithm") visa automatizar a divisão de conjuntos de dados de moléculas em conjuntos de treinamento e teste para modelagem QSAR/QSPR. Utilizando informações sobre a estrutura da molécula, propriedades físico-químicas e atividades biológicas para realizar uma amostragem racional e representativa do espaço químico, visando gerar modelos mais robustos e preditivos, evitando vícios e otimizando a cobertura do espaço químico durante a construção do modelo. Para torná-lo acessível a um público mais amplo, está sendo desenvolvido uma interface de usuário amigável em formato de aplicação web. Hospedada em um servidor da UFMG, a aplicação permitirá que usuários sem expertise em computação possam facilmente carregar seus conjuntos de dados, configurar parâmetros e visualizar os resultados de forma clara e concisa.

## 1. Introdução

A modelagem molecular tem se mostrado uma ferramenta poderosa na descoberta e desenvolvimento de novos fármacos, auxiliando na identificação de compostos líderes e na otimização de candidatos promissores. Dentre as diversas técnicas existentes, a modelagem QSAR/QSPR (Quantitative Structure-Activity/Property Relationship) [Todeschini and Consonni 2009] destaca-se por correlacionar a estrutura química de uma molécula à sua atividade biológica ou propriedade físico-química. Essa abordagem permite a triagem virtual de bibliotecas de compostos e a predição da atividade de novas moléculas, acelerando o processo de desenvolvimento de fármacos e reduzindo custos.

A robustez e confiabilidade dos modelos QSAR/QSPR são diretamente influenciadas pela qualidade dos dados utilizados em sua construção, especialmente na etapa de divisão do conjunto de dados em conjuntos de treinamento e teste. A seleção aleatória, embora amplamente utilizada, pode resultar em modelos com baixa capacidade preditiva e generalização limitada, especialmente em conjuntos de dados com alto grau de variabilidade estrutural e/ou com número reduzido de moléculas. Para contornar essa limitação, métodos de amostragem racional, como Kennard-Stone [Ferreira et al. 2022] e SPXY, têm sido propostos, buscando garantir uma divisão mais representativa do espaço químico.

No entanto, a maioria dos métodos existentes não leva em consideração simultaneamente a informação contida na estrutura da molécula, propriedades físico-

químicas e atividades biológicas durante a divisão dos dados. Essa lacuna motivou o desenvolvimento do algoritmo MASSA ("Molecular dAta Set SAmpling Algorithm") [Veríssimo et al. 2023], que busca suprir as limitações das abordagens tradicionais, explorando a sinergia entre diferentes espaços químicos para realizar uma amostragem mais eficiente e gerar modelos QSAR/QSPR mais robustos e preditivos.

Visando ampliar o alcance e facilitar o uso do algoritmo MASSA pela comunidade científica, este trabalho propõe torná-lo mais acessível através de uma aplicação web com interface amigável e intuitiva, hospedada em um servidor da UFMG. A fim de garantir a qualidade e manutenibilidade do código a longo prazo, o código do algoritmo será também refatorado utilizando os princípios do clean code [Martin 2009] e as diretrizes do PEP 8 [Van Rossum et al. 2001], tornando-o mais legível e modular.

### 1.1. Objetivo Geral

Desenvolver um software para democratizar e facilitar o uso do algoritmo MASSA, permitindo a automatização da divisão de conjuntos de dados de moléculas para modelagem QSAR/QSPR.

### 1.2. Objetivos Específicos

A seguir seguem os objetivos específicos:

- Refatoração: Tornar o código mais limpo, coeso e flexível usando princípios SOLID e clean code, facilitando a manutenção e expansão.
- Padronização: Seguir PEP 8 para melhor legibilidade e documentação clara com Swagger.
- Interface amigável: Criar uma interface intuitiva com VUEjs, design limpo, navegação fácil e widgets adequados.
- Relatórios versáteis: Permitir geração de relatórios em PDF com visualizações gráficas e opções de personalização.
- Tutorial completo: Integrar um tutorial passo a passo que facilite o entendimento do usuário.
- Coleta de dados anônimos: Coletar dados de uso anônimos com Google Analytics para direcionar melhorias na usabilidade.
- Armazenamento seguro: Proteger o algoritmo e os dados com criptografia e seguir as políticas de segurança da UFMG.

### 1.3. Justificativa do Projeto

A demanda por inovação na descoberta de fármacos exige ferramentas computacionais eficientes, como os modelos QSAR/QSPR, que preveem propriedades moleculares e otimizam o processo. No entanto, a eficácia desses modelos depende da qualidade da amostragem de dados, frequentemente limitada por abordagens aleatórias que comprometem a capacidade preditiva.

O algoritmo MASSA ("Molecular dAta Set SAmpling Algorithm") surge como uma solução inovadora para essa limitação. Integrando dados biológicos, físico-químicos e estruturais, o MASSA oferece uma amostragem mais sofisticada, resultando em modelos QSAR/QSPR mais robustos e confiáveis.

Sua disponibilidade em uma aplicação web acessível expande o alcance da ferramenta, permitindo que pesquisadores sem expertise em programação se beneficiem de técnicas avançadas de amostragem. Essa democratização do acesso impulsiona a descoberta de fármacos e beneficia a saúde pública e a ciência como um todo.

Em suma, o MASSA otimiza a modelagem QSAR/QSPR, tornando a triagem de compostos mais eficiente e precisa. Sua interface amigável e acessibilidade promovem a disseminação de tecnologias inovadoras, impulsionando avanços significativos na química computacional e tornando a modelagem molecular acessível à comunidade científica global.

## 2. Referencial Teórico

As técnicas de modelagem computacional, como QSAR (Quantitative Structure-Activity Relationship) e QSPR (Quantitative Structure-Property Relationship), são amplamente utilizadas para prever propriedades biológicas e físico-químicas de compostos químicos a partir de suas estruturas moleculares. Essas técnicas permitem correlacionar as características estruturais das moléculas com suas atividades biológicas ou propriedades, desempenhando um papel fundamental na descoberta de novos fármacos (Todeschini e Consonni, 2009).

Diversos métodos têm sido desenvolvidos para construir modelos preditivos robustos, incluindo técnicas de aprendizado de máquina, como regressão linear e redes neurais (Cherkasov et al., 2014). A eficácia desses modelos, no entanto, depende em grande parte da qualidade dos dados e de como os conjuntos de treinamento e teste são definidos.

A amostragem de dados é um passo crítico nesse processo, pois garante a representatividade do espaço químico nos modelos. Abordagens tradicionais, como a divisão aleatória, podem não cobrir adequadamente todas as regiões do espaço químico, resultando em baixa generalização dos modelos (Golbraikh e Tropsha, 2002). Métodos racionais, como o Kennard-Stone, buscam resolver essa questão, mas ainda apresentam limitações ao não integrarem completamente as diferentes dimensões dos dados, como propriedades biológicas e físico-químicas (Kennard e Stone, 1969).

Nos últimos anos, ferramentas computacionais como KNIME e R têm facilitado o desenvolvimento de modelos QSAR/QSPR, mas muitas delas exigem conhecimento técnico especializado e não integram soluções completas para a amostragem de dados (Berthold et al., 2008; R Core Team, 2021). Foi nesse contexto que surgiu o algoritmo MASSA, uma abordagem automatizada para divisão de conjuntos de dados, que utiliza técnicas como PCA, HCA e K-modes.

O MASSA ("Molecular dAta Set SAmpling Algorithm") automatiza a divisão de dados moleculares em conjuntos de treinamento e teste, abordando as limitações dos métodos tradicionais. Ele integra de forma sinérgica as dimensões biológica, físico-química e estrutural das moléculas, garantindo uma amostragem representativa e equilibrada. A utilização de técnicas de análise multivariada permite a criação de modelos mais robustos, com melhores métricas de validação.

Os resultados obtidos com o MASSA mostram uma melhora significativa na construção de modelos preditivos. A interface web amigável que está sendo desenvolvida para o algoritmo também garante que ele seja acessível a um público mais amplo,

incluindo usuários sem conhecimentos técnicos avançados. Essa inovação contribui diretamente para a racionalização do processo de amostragem de dados, promovendo a construção de modelos mais confiáveis e precisos.

## **2.1. Extensão Universitária**

A extensão universitária é um trabalho que transcende os muros da academia, estabelecendo um elo dinâmico entre a universidade e a sociedade. É um processo educativo, cultural e social que promove a troca de saberes e experiências entre a comunidade acadêmica e os diferentes segmentos da sociedade. Através de ações e projetos que dialogam com as necessidades e demandas do contexto social, a extensão universitária contribui para a democratização do conhecimento, a transformação social e a busca por soluções para os desafios contemporâneos.

Este projeto de desenvolvimento do software para o uso do algoritmo MASSA representa uma iniciativa de extensão universitária que conecta o conhecimento acadêmico com a realidade prática, beneficiando tanto a comunidade acadêmica quanto a sociedade em geral. Ao democratizar o acesso a ferramentas avançadas de modelagem, como o algoritmo MASSA, estamos desenvolvendo habilidades essenciais entre os pesquisadores, especialmente no uso de tecnologias computacionais em suas pesquisas.

A colaboração com o Laboratório de Farmácia da UFMG aproxima a universidade da sociedade, oferecendo uma solução prática para um problema real enfrentado por pesquisadores que não possuem familiaridade com linhas de comando. Ao criar uma interface amigável e intuitiva para o algoritmo, o projeto facilita o acesso a tecnologias avançadas, ampliando o impacto das pesquisas realizadas na universidade e promovendo uma maior inclusão na aplicação de métodos computacionais em diferentes áreas da ciência.

Além disso, o projeto forma agentes de mudança ao capacitar estudantes e pesquisadores para o uso de novas ferramentas e tecnologias, preparando-os para contribuir de forma significativa em seus campos de atuação. Este esforço está alinhado com o Objetivo de Desenvolvimento Sustentável (ODS) 3: "Saúde e Bem-estar", pois ao otimizar as técnicas de modelagem QSAR/QSPR, contribuímos para avanços na pesquisa farmacêutica e, consequentemente, para a melhoria da saúde e bem-estar da sociedade.

Este projeto também é parte do esforço mais amplo do Núcleo de Extensão Universitária, que apoia e promove iniciativas que conectam a universidade com a comunidade, fortalecendo o papel da instituição como agente de transformação social. Mais informações sobre o Núcleo de Extensão Universitária e suas iniciativas estão disponíveis em [www.pucminas.br/proex](http://www.pucminas.br/proex).

## **2.2. Parceiro**

O Laboratório de Farmácia da UFMG, é o parceiro principal deste projeto. Esta colaboração surge com o objetivo de democratizar o uso de ferramentas avançadas de modelagem, como o algoritmo MASSA. O laboratório possui um foco significativo em inovação e eficiência no desenvolvimento de modelos QSAR/QSPR, que são fundamentais para a previsão de atividades biológicas de compostos químicos.

Atualmente, o algoritmo MASSA está disponível apenas via linhas de comando, o que limita seu acesso a pesquisadores com algum conhecimento técnico em programação.

Buscando democratizar o acesso a essa ferramenta inovadora e torná-la acessível e utilizável por todos os pesquisadores, independente de suas habilidades de programação, essa parceria com o laboratório surge como um esforço conjunto para superar barreiras tecnológicas e promover a inclusão na aplicação de ferramentas computacionais em pesquisas farmacêuticas.

A colaboração com o Laboratório de Farmácia da UFMG é fundamental para garantir que o algoritmo MASSA possa ser utilizado de forma mais ampla e eficiente. Ao criar uma interface amigável e intuitiva, o projeto visa reduzir as dificuldades associadas ao uso de ferramentas que operam apenas por linha de comando, encorajando mais pesquisadores a adotarem essas tecnologias avançadas em seus trabalhos. O laboratório, através dessa parceria, reforça seu compromisso com a inovação e a acessibilidade, buscando sempre facilitar o trabalho dos seus pesquisadores com soluções tecnológicas eficientes.

### **2.3. Trabalhos relacionados**

A modelagem QSAR (Quantitative Structure–Activity Relationship) é uma abordagem amplamente utilizada na descoberta de fármacos para prever a atividade biológica de compostos químicos com base em suas estruturas moleculares. Vários trabalhos importantes têm contribuído para o avanço dessa área de pesquisa, destacando diferentes métodos e abordagens.

O trabalho de Todeschini e Consonni (2009) em "QSAR Models for Prediction of Biological Activity: Methods and Techniques" é um dos marcos na aplicação de métodos matemáticos e estatísticos para a construção de modelos QSAR. Este livro fornece uma base teórica abrangente, abordando a seleção de descritores moleculares, técnicas de validação cruzada e métodos para evitar o sobreajuste. A obra enfatiza a importância de um processo rigoroso de desenvolvimento e validação de modelos para garantir que os modelos QSAR sejam robustos e generalizáveis, capazes de prever atividades biológicas de novos compostos de maneira confiável.

Cherkasov et al. (2014), em seu artigo "Quantitative Structure–Activity Relationships (QSAR) in Drug Discovery", publicado na Chemical Reviews, complementa essa visão fornecendo uma análise abrangente do uso de QSAR na descoberta de fármacos. Eles discutem não apenas os fundamentos da modelagem QSAR, mas também as aplicações práticas e os desafios enfrentados ao utilizar esses modelos no desenvolvimento de novos medicamentos. Cherkasov et al. destacam a integração de técnicas computacionais avançadas e a necessidade de métodos eficientes para a seleção de conjuntos de treinamento e teste, que são essenciais para a construção de modelos preditivos confiáveis.

O artigo de Veríssimo et al. (2023) apresenta o algoritmo MASSA (Molecular dAta Set SAMpling Algorithm), que se propõe a melhorar o processo de divisão de conjuntos de dados de moléculas em treinamento e teste para modelagem QSAR/QSPR (Quantitative Structure-Property Relationship). Ao contrário das abordagens tradicionais que dependem frequentemente de seleções aleatórias ou arbitrárias de dados, o algoritmo MASSA utiliza informações estruturais das moléculas, propriedades físico-químicas e atividades biológicas para realizar uma amostragem racional e representativa do espaço químico. Essa abordagem visa a criação de conjuntos de dados mais equilibrados, minimizando vieses e otimizando a cobertura do espaço químico, o que resulta em modelos

preditivos mais robustos.

Enquanto Todeschini e Consonni fornecem a base teórica e metodológica, e Cherkasov et al. exploram aplicações e desafios práticos no contexto da descoberta de fármacos, o trabalho de Veríssimo et al. foca especificamente na otimização do processo de preparação de dados para modelagem QSAR. A contribuição inovadora do algoritmo MASSA está em seu método sistemático de divisão de dados, que evita problemas comuns de sobreajuste e garante que os modelos QSAR sejam treinados e testados em conjuntos de dados que são representativos e equilibrados.

Assim, o artigo de Veríssimo et al. complementa os trabalhos de Todeschini e Consonni e Cherkasov et al. ao introduzir uma abordagem prática e automatizada para a seleção de conjuntos de dados. Essa abordagem não só melhora a qualidade dos modelos QSAR desenvolvidos, mas também aborda diretamente alguns dos desafios destacados na literatura existente, como a necessidade de métodos mais eficientes e representativos para a divisão de conjuntos de dados. Portanto, o algoritmo MASSA representa um avanço significativo no campo da modelagem QSAR, proporcionando uma ferramenta que pode ser integrada com as técnicas estabelecidas para melhorar ainda mais a precisão e a aplicabilidade dos modelos preditivos.

### **3. Metodologia**

Este artigo detalha a metodologia empregada no desenvolvimento do software para democratizar o algoritmo MASSA, uma inovadora solução para amostragem de dados moleculares com o objetivo de aprimorar a qualidade e o poder preditivo de modelos QSAR/QSPR. A abordagem escolhida foi o framework Scrum, permitindo um desenvolvimento ágil e iterativo, dividido em cinco sprints. Cada sprint focou na entrega incremental de funcionalidades, priorizando valor para o usuário e permitindo flexibilidade para incorporar feedback e adaptar-se a mudanças. A seguir, descrevemos a estrutura de cada sprint, detalhando as decisões de, o desenvolvimento da interface web, as escolhas arquiteturais, a implementação do back-end e outros aspectos relevantes.

#### **3.1. Sprint 1**

A primeira sprint focou em estabelecer uma base sólida para o desenvolvimento do software. O objetivo principal foi compreender o problema e definir o escopo do projeto. Para isso, foram realizadas reuniões com o Prof. Vinicius e o Dr. Gabriel Corrêa Veríssimo, criador do algoritmo MASSA, para entender as dificuldades do uso atual e as funcionalidades desejadas para torná-lo mais acessível. Essa etapa de levantamento de requisitos foi crucial para contextualizar o problema, definindo claramente o público-alvo e os objetivos da interface web. Com base nessas informações, foi formalizado um acordo com o cliente, estabelecendo o escopo do projeto, prazos e entregas, alinhados com as necessidades dos stakeholders. Finalmente, o problema e a solução proposta foram apresentados para toda a equipe de desenvolvimento, garantindo uma visão compartilhada e um entendimento comum dos objetivos da sprint.

#### **3.2. Sprint 2**

Com o problema bem definido na Sprint 1, a Sprint 2 concentrou-se na modelagem da solução e na criação de um protótipo para validação do design e do fluxo de usuário.

Diagramas de caso de uso e um diagrama ER foram criados para representar visualmente o funcionamento do software e as interações do usuário. Um protótipo interativo das telas da interface web foi desenvolvido utilizando ferramentas de prototipagem, permitindo a demonstração do fluxo de usuário e a coleta de feedback. A jornada do usuário dentro do software foi detalhadamente definida, identificando as principais ações e interações com as funcionalidades. Por fim, a estrutura básica das páginas principais da interface web, como a página inicial e a página de login, foi implementada utilizando HTML, CSS e JavaScript.

### **3.3. Sprint 3**

A Sprint 3 teve como foco o refinamento da solução, a implementação das funcionalidades principais do software e a criação de uma versão inicial da API. Os diagramas de caso de uso e o diagrama ER foram revisados e aprimorados com base no feedback recebido sobre o protótipo e nas novas funcionalidades identificadas. O diagrama ER foi traduzido para um modelo de dados relacional, definindo as tabelas e relações do banco de dados, que foi posteriormente otimizado para garantir eficiência e integridade das informações. A equipe dedicou-se à implementação do código back-end das funcionalidades principais, incluindo a integração com o algoritmo MASSA, a página "Sobre", e a página de suporte. Finalmente, uma versão inicial da API foi criada para a comunicação entre o front-end e o back-end, permitindo o acesso aos dados e funcionalidades do software, juntamente com a implementação do sistema de armazenamento de dados.

### **3.4. Sprint 4**

A Sprint 4 iniciou-se com a validação do novo layout do front-end pelo cliente, assegurando que atendesse às suas expectativas e necessidades. Em seguida, o foco principal foi a integração entre front-end e back-end, permitindo a execução do algoritmo MASSA diretamente pela interface. Para viabilizar a interação com os dados plotados, o algoritmo foi modificado para retornar dados estruturados em vez de imagens, possibilitando a implementação de gráficos interativos no front-end e uma visualização mais clara dos resultados. O desenvolvimento da exportação para PDF foi concluído no back-end, restando apenas a integração com o front-end. Visando atender às necessidades do projeto, a execução do algoritmo foi dividida em duas etapas: análise do arquivo enviado pelo usuário e execução com os parâmetros extraídos e selecionados. A Sprint 4 também incluiu a refatoração da análise de atividades biológicas, a implementação de sessões de usuário e a revisão e o aprimoramento dos diagramas de caso de uso e do diagrama ER, incorporando o feedback e as novas funcionalidades. Além disso, foram implementados testes de integração e unidade para garantir a qualidade e a robustez das novas funcionalidades.

### **3.5. Sprint 5**

A Sprint 5 focou na implementação de funcionalidades adicionais, como a descrição detalhada dos parâmetros extraídos e a inclusão de títulos para cada bloco que os contém. O sistema de login foi completamente integrado, com validações de segurança aprimoradas utilizando tokens. A página de "Dados", que exibe informações como o número médio de moléculas por atividade biológica, o total de atividades biológicas analisadas e a localização dos usuários (identificada por meio do IP), foi totalmente integrada e validada com o backend. Textos placeholders, como "Lorem Ipsum", foram substituídos por

conteúdos definitivos validados com o cliente, e o deploy completo da aplicação foi realizado. O backend e o banco de dados foram hospedados na plataforma Heroku, enquanto o frontend foi feito o deploy na plataforma Netlify.

## **4. Resultados**

Esta seção apresenta o resultado consolidado do projeto, incluindo a arquitetura final do software, seus componentes, artefatos definitivos (diagrama de classes, modelo de dados relacional, lista de requisitos), telas da aplicação, bem como as integrações e funcionalidades tal como entregues na versão final. Aqui não serão mostradas versões preliminares ou protótipos, mas apenas o produto final após todos os refinamentos e melhorias realizadas ao longo do desenvolvimento.

### **4.1. Descrição Técnica da Solução**

O MASSA Web oferece uma interface gráfica intuitiva e acessível para o MASSA Algorithm, um software desenvolvido para a divisão de datasets em conjuntos de treino e teste, muito útil para a construção de modelos de aprendizado de máquina e estudos de QSAR. Além disso, o MASSA Algorithm demonstrou superioridade na qualidade dos modelos gerados em comparação com métodos de amostragem tradicionalmente empregados na área de desenvolvimento de fármacos. Com o MASSA Web, os usuários podem realizar a amostragem de datasets de forma prática e a partir de qualquer lugar, eliminando a necessidade de lidar com linhas de comando ou instalar interpretadores Python localmente.

Além disso, a interface gráfica proporciona maior controle e interatividade na análise de dados, permitindo o ajuste dinâmico de gráficos, com ferramentas de zoom, manipulação de eixos e visualização detalhada das distribuições químicas. Essa abordagem não só simplifica o processo, mas também aumenta a produtividade e acessibilidade, permitindo o acesso também para usuários iniciantes que não queiram lidar com a linha de comando.

### **4.2. Arquitetura do Software**

#### **4.2.1. Backend**

A arquitetura do backend do sistema segue o modelo de arquitetura em camadas, uma abordagem amplamente utilizada no desenvolvimento de software para organizar aplicações em módulos separados por responsabilidades, promovendo modularidade e escalabilidade.

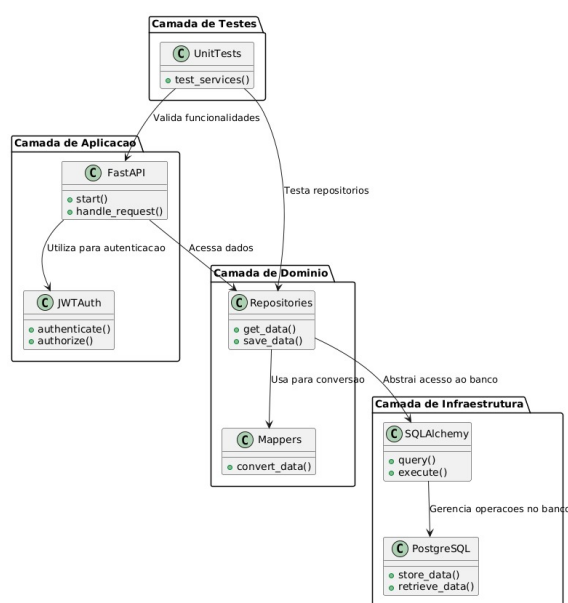
Para a implementação, utilizamos o FastAPI, um framework Python moderno e de alta performance, ideal para construir APIs robustas e rápidas. Como camada de abstração entre a aplicação e o banco de dados relacional (PostgreSQL), optamos pelo SQLAlchemy, um ORM (Object-Relational Mapping) que facilita a manipulação de dados de forma orientada a objetos, mantendo a flexibilidade necessária para realizar consultas SQL personalizadas.

Além disso, a arquitetura do backend é composta por repositories, que centralizam as operações de acesso ao banco de dados, promovendo o desacoplamento entre as camadas. A segurança é garantida por meio de uma camada de autenticação e autorização baseada em JWT (JSON Web Tokens), que protege as rotas e os dados transmitidos.



Outros elementos importantes incluem o uso de Schemas, que estruturam e validam os dados trafegados, Helpers, que encapsulam funcionalidades utilitárias reutilizáveis, e Mappers, responsáveis por realizar a conversão entre diferentes objetos e formatos dentro do sistema, promovendo consistência e organização.

Por fim, o backend do sistema conta com uma camada de testes robusta, composta por testes unitários para todas as classes de serviço. Essa abordagem garante a validação individual de cada funcionalidade, contribuindo para a prevenção de falhas e para a manutenção contínua do sistema, que se torna mais confiável e escalável ao longo do tempo. A Figura 1 apresenta o diagrama de pacotes do nosso sistema. Para mais detalhes, a imagem pode ser acessada [aqui](#).



**Figura 1. Diagrama de Pacotes do Sistema**

#### 4.2.2. Frontend

O frontend do sistema foi desenvolvido utilizando Vue.js, uma estrutura moderna baseada em JavaScript, com o suporte do Vuetify, uma biblioteca baseada em Material Design que oferece uma interface consistente e responsiva. A estrutura do projeto segue uma organização modular, incluindo pastas como assets para arquivos estáticos, components para componentes reutilizáveis, layouts para estrutura da interface, router para gerenciamento de rotas, services para comunicação com a API, stores para gerenciamento de estado e views para as páginas principais da aplicação.

#### 4.3. Artefatos

Os artefatos utilizados no desenvolvimento do sistema garantiram organização, planejamento e comunicação eficiente durante todo o processo. O levantamento de requisitos foi realizado diretamente com o cliente, acompanhado por reuniões frequentes para feedback contínuo ao longo do desenvolvimento.

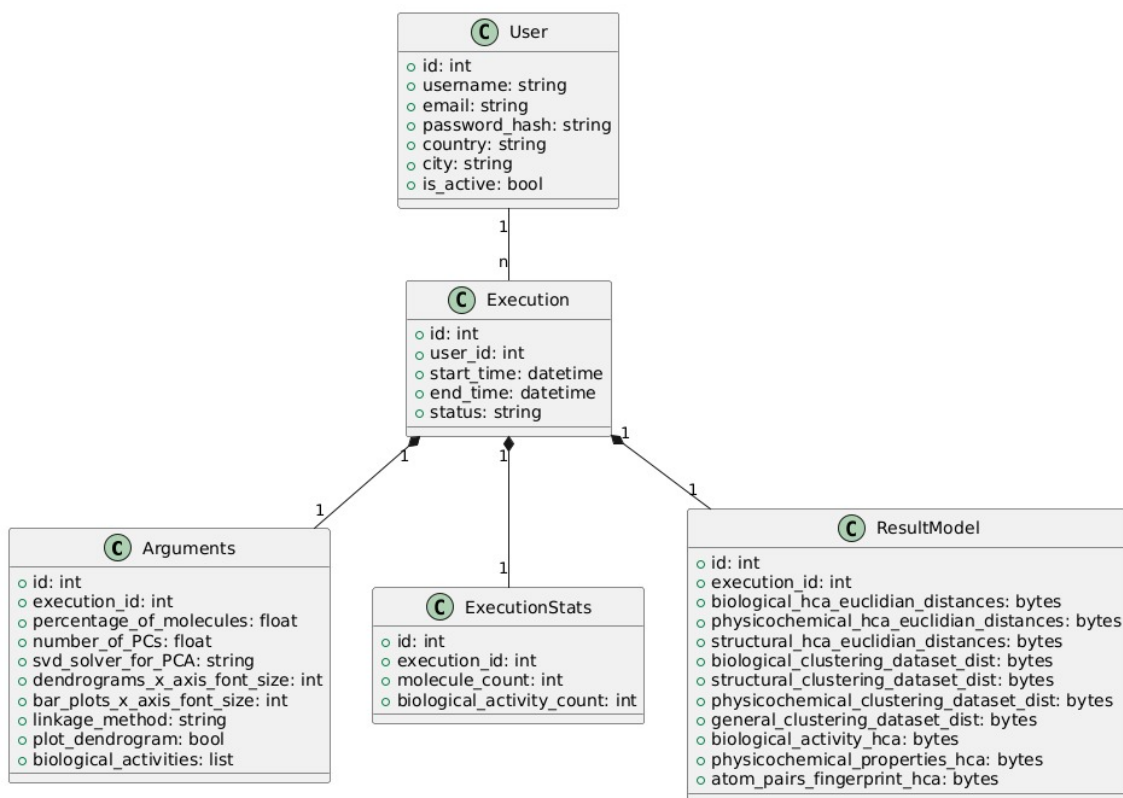
A organização e o planejamento das sprints foram conduzidos utilizando o kanban, permitindo uma visualização clara das tarefas e do progresso. Adotamos um padrão de desenvolvimento organizado, no qual cada issue é desenvolvida em uma branch específica. Após a conclusão, é criado um pull request (PR) para a branch de desenvolvimento, e outro membro da equipe realiza a aprovação ou rejeição do PR. Além disso, foi estabelecido um padrão de commits para garantir rastreabilidade e clareza no histórico do repositório.

Para a interface do sistema, desenvolvemos um protótipo do layout utilizando o Figma. A modelagem técnica foi documentada com diversos diagramas: diagrama de componente, diagrama de container, diagrama entidade-relacionamento, diagrama de contexto do sistema, diagrama de caso de uso e diagrama de classe. Esses artefatos auxiliaram no planejamento, entendimento e comunicação clara entre os membros da equipe e os stakeholders, garantindo a qualidade e alinhamento do sistema desenvolvido.

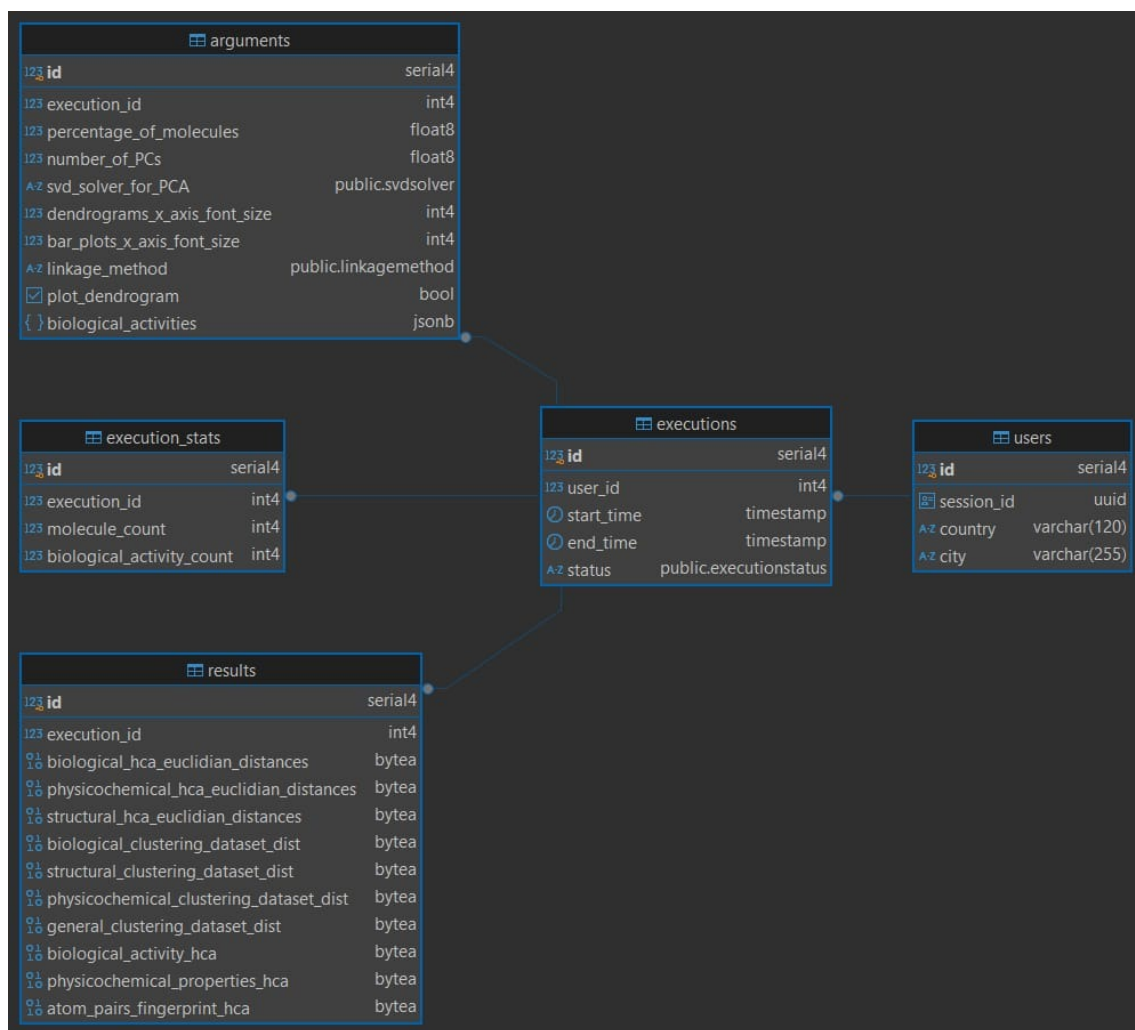
A Figura 2 apresenta o diagrama de classes do nosso sistema. Para mais detalhes, a imagem pode ser acessada [aqui](#).

A Figura 3 apresenta o modelo relacional do nosso sistema.

A Tabela 1 representa os Requisitos Funcionais, Tabela 2, requisitos não funcionais do sistema



**Figura 2. Diagrama de Classes do Sistema**



**Figura 3. Modelo Relacional do Sistema**

<b>Número de Ordem</b>	<b>Requisito</b>	<b>Descrição</b>	<b>Prioridade</b>
RF01	O software deve gerar uma página de reports detalhados sobre o processamento do MASSA nas entradas de dados do usuário.	Após o processamento das entradas do usuário, o software deve detalhar as saídas em uma nova página, por meio de dendrogramas, gráficos e outros recursos.	Alta
RF02	O software deve permitir que o usuário baixe os reports gerados em formato PDF.	Os relatórios gerados pelo software devem poder ser baixados em formato PDF, garantindo portabilidade e compatibilidade com diferentes sistemas.	Alta
RF03	O software deve salvar os dados recentes da sessão do usuário.	O software deve armazenar os dados de entrada, configurações utilizadas e resultados da sessão atual do usuário (apenas browser), permitindo que ele acesse essas informações posteriormente.	Baixa
RF04	O software deve possuir um tutorial inicial disponível para o usuário.	Um conjunto de instruções claras e concisas devem estar disponíveis para auxiliar os usuários iniciantes a utilizarem as funcionalidades básicas do software.	Média
RF05	O software deve possuir uma aba específica, com um dashboard de configuração dos parâmetros do algoritmo.	O software deve oferecer uma interface gráfica que permite ao usuário configurar os parâmetros do algoritmo, como opções de processamento e formato de saída.	Alta
RF06	O software deve coletar dados de métricas de uso definidas sobre sua utilização.	O software coletará dados a fim de gerar dados anonimizados sobre o uso do software, como número de usuários, funcionalidades mais utilizadas e erros reportados, que auxiliem na tomada de decisões futuras sobre ele.	Média
RF07	Refatoração do algoritmo do MASSA para um código mais limpo e legível.	Refatorar o algoritmo do MASSA, para permitir que o código-fonte do algoritmo tenha leitura e manutenção fáceis por interessados em utilizá-lo em pesquisas.	Baixa

**Tabela 1. Requisitos Funcionais**

Número de Ordem	Requisito	Descrição	Prioridade
RNF01	A interface do usuário deve ser amigável e fácil de usar.	A interface do usuário deve ser intuitiva e de fácil navegação, permitindo que usuários sem experiência prévia consigam utilizar o software sem dificuldades.	Alta
RNF02	A interface do usuário deve ser limpa e intuitiva, permitindo que o usuário navegue facilmente pelo software e analise os resultados de forma eficaz.	O design da interface deve ser limpo e organizado, com elementos visuais claros e informações apresentadas de forma concisa, facilitando a compreensão e análise dos resultados.	Alta
RNF03	Os reports em PDF gerados pelo software devem ser padronizados com informações do software e do laboratório, incluindo a identidade visual.	Os relatórios em PDF devem seguir um template padrão contendo nome e versão do software, nome e contato do laboratório, data e hora da geração do relatório, além da identidade visual (ex: logotipo) previamente definida.	Alta
RNF04	O software deve ser protegido contra múltiplas requisições ao mesmo tempo (Rate Limit).	O software deve ser capaz de lidar com múltiplas requisições simultâneas sem comprometer a performance ou a integridade dos dados, garantindo o funcionamento adequado em situações de alta demanda.	Média
RNF05	O software deve garantir disponibilidade aos usuários na maior parte do tempo.	O software deve estar acessível aos usuários ininterruptamente, idealmente com mecanismos de tolerância a falhas para minimizar o tempo de inatividade em caso de eventuais problemas técnicos.	Baixa

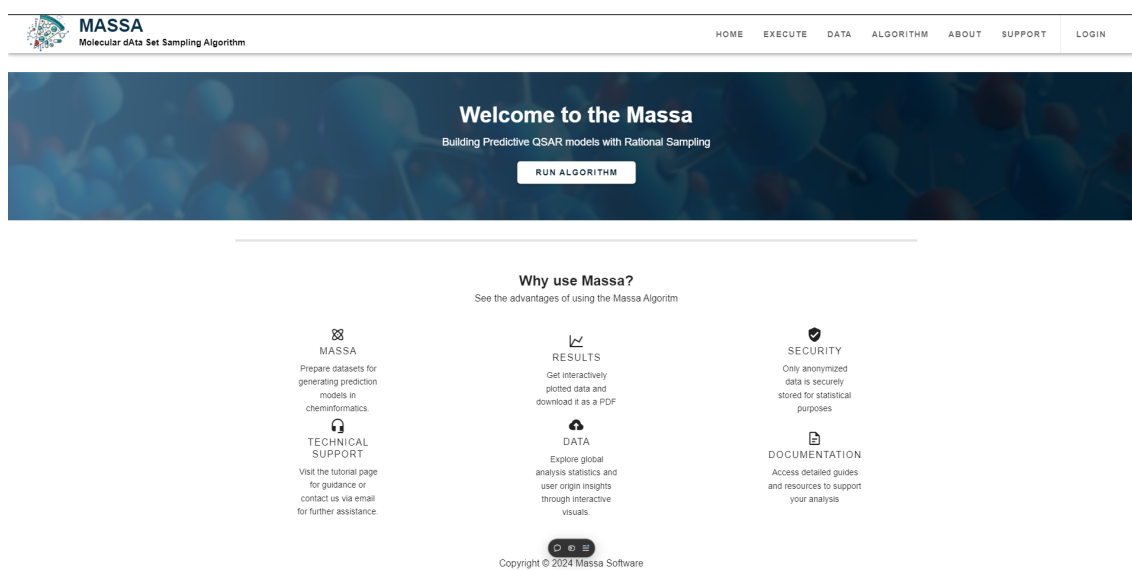
**Tabela 2. Requisitos Não Funcionais**

#### **4.4. Telas da aplicação**

A Figura 4 apresenta a tela de página inicial da aplicação, que oferece um botão para a execução do algoritmo MASSA. A página também possui um menu de navegação com links para outras seções (Home, Execute, Data, Algorithm, About, Support e Login). Abaixo da seção principal, há uma área intitulada "Why use Massa?", que destaca as

vantagens da utilização do algoritmo e da plataforma.

Além das informações visuais na tela, a descrição fornecida detalha as funcionalidades e o funcionamento do Projeto MASSA. A seção "How the algorithm works" explica a metodologia do algoritmo, que utiliza PCA, HCA e K-modes para garantir uma representação equilibrada da diversidade do conjunto de dados nos conjuntos de treinamento e teste. A seção "Upload your data" descreve os formatos de arquivo suportados para entrada de dados, com preferência pelo formato SDF. A seção "Smart Sampling" detalha o processo de amostragem, que envolve o cálculo de propriedades físico-químicas e estruturais, clustering independente dos domínios e clustering K-modes para unificar a divisão dos clusters.



**Figura 4. Application Home Page**

A Figura 5 apresenta a tela de execução do algoritmo MASSA, divididos em três seções principais: "Upload File", "Optional Arguments" e "Results and download". A seção "Upload File", mostrada na Figura 6 permite ao usuário carregar o conjunto de dados moleculares para análise. A seção "Optional Arguments", mostrada na Figura 6 contém os parâmetros ajustáveis pelo usuário, enquanto a seção "Results and download", mostrada na Figura 7, é utilizada posteriormente para apresentar os resultados e permitir o download destes.

Dentro de "Optional Arguments", o primeiro parâmetro é "Percentage of molecules in training set", controlado por um slider que define a porcentagem de moléculas alocadas ao conjunto de treinamento. O valor padrão é , e a porcentagem para o conjunto de teste é calculada como o complemento. O campo "Biological activities for separation" permite a especificação das atividades biológicas ou outras propriedades a serem consideradas na modelagem QSAR ou de aprendizado de máquina. O algoritmo suporta múltiplas propriedades, desde que representadas como números inteiros ou de ponto flutuante. Classes devem ser codificadas numericamente. "Number of principal components in PCA" define o número de componentes principais para a redução de dimensionalidade.

Valores decimais entre 0 e 1 representam a porcentagem de variância explicada, enquanto valores inteiros maiores que 1 especificam o número exato de PCs. Há regras específicas para o número de PCs em relação ao número de propriedades físico-químicas e atividades biológicas. "SVD solver for PCA" está fixado em "full" na versão web. "HCA linkage method" permite a escolha do critério de ligação para o clustering hierárquico. "X-axis font for Dendrograms" permite configurar o tamanho da fonte do eixo X dos dendrogramas. Por fim, "Plot dendrogram" é um interruptor que define se o dendrograma será gerado.

The screenshot shows the MASSA application interface. The header includes the MASSA logo and the text "Molecular dAta Set Sampling Algorithm". The navigation bar contains links: HOME, EXECUTE, DATA, ALGORITHM, ABOUT, SUPPORT, and LOGIN. The main content area is divided into three steps: 1. Upload file, 2. Optional Arguments, and 3. Results and download. The "Upload file" step is active, showing a large "UPLOAD FILE" button. Below the button are "PREVIOUS" and "NEXT" navigation links. At the bottom, there is a copyright notice: "Copyright © 2024 Massa Software".

Figura 5. Application Execution Page

The screenshot shows the MASSA application interface, specifically the "Optional Arguments" step. The header and navigation bar are the same as in Figure 5. The main content area is divided into three steps: 1. Upload file, 2. Optional Arguments, and 3. Results and download. The "Optional Arguments" step is active, showing various configuration options. These include a slider for "Percentage of molecules in training set" set to 80%, a dropdown for "Biological activities for separation", input fields for "N° of principal components in PCA" (0,85), "SVD solver for PCA" (full), and "HCA linkage method" (complete). Below these are input fields for "X-axis font for Dendrograms" (5) and "X-axis font for Bar Charts" (12), and a toggle switch for "Plot dendrogram" which is currently turned on. At the bottom, there are "PREVIOUS" and "NEXT" navigation links, and a copyright notice: "Copyright © 2024 Massa Software".

Figura 6. Application Execution Page 2



**Figura 7. Application Execution Page 3**

O Software MASSA oferece telas de registro Figura 8 e login Figura 9 simples e intuitivas para gerenciamento de usuários. A tela de registro coleta informações essenciais para a criação de uma nova conta, incluindo nome de usuário, endereço de email, senha e confirmação de senha. Campos de senha incluem validação de quantidade mínima de caracteres e o campo de e-mail possui validações para que o e-mail fornecido seja válido. Botões "Register" e "Cancel" permitem que o usuário prossiga com o cadastro ou cancele a operação.

A tela de login, por sua vez, foca na autenticação do usuário, solicitando apenas nome de usuário e senha. Similarmente à tela de registro, o campo de senha possui validação de quantidade mínima de caracteres. Além do botão "Login" para acessar a aplicação, a tela também oferece um botão "Register" para direcionar novos usuários ao formulário de cadastro. Ambas as telas mantêm a identidade visual da aplicação, com cabeçalho, menu de navegação e rodapé padronizados, garantindo uma experiência de usuário consistente e coesa.




## Register in Massa Software

### Username

 Insert your username

### Email

 Insert your email

### Password

Insert your password



Confirm your password




Register

Cancel

Figura 8. Application Register Page


## Login in Massa Software

**Username**

 Insert your username

**Password**

Insert your password



[Forgot login password?](#)

Login

Register

**Figura 9. Application Login Page**

A página de estatísticas do usuário do software MASSA oferece uma visão geral do uso da plataforma, apresentando dados agregados sobre as análises realizadas pelo algoritmo. A página é organizada em duas abas: "MOLECULES COUNT" e "CITY AND COUNTRY INFOS". Na aba "MOLECULES COUNT", mostrada na Figura 10, o usuário encontra estatísticas sobre as atividades biológicas analisadas, como o número total de atividades, a média de moléculas por atividade e o total de análises realizadas até o momento. Um gráfico de linhas ilustra a distribuição das análises ao longo da semana, permitindo visualizar padrões de uso.

A aba "CITY AND COUNTRY INFOS", mostrada na Figura ?? concentra-se na localização geográfica dos usuários. Uma tabela exhibe as cidades e países de origem dos usuários que utilizaram o Software. Uma barra de pesquisa facilita a busca por cidades específicas, e botões de paginação permitem navegar por uma lista extensa de localidades. Em ambas as abas, o layout consistente com o restante da aplicação, incluindo cabeçalho, menu de navegação e rodapé, garante uma experiência de usuário coesa e intuitiva. Es-

sas estatísticas fornecem informações valiosas sobre o alcance e o impacto do Software MASSA, permitindo aos desenvolvedores e usuários compreenderem melhor o perfil da comunidade e as tendências de utilização da ferramenta.

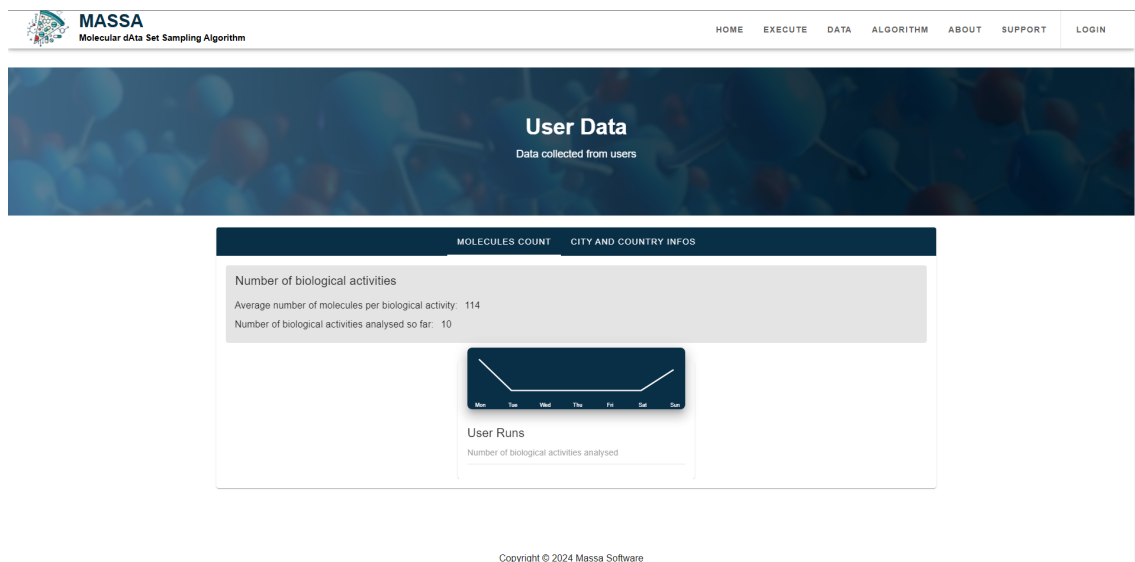


Figura 10. Application Run Stats Page

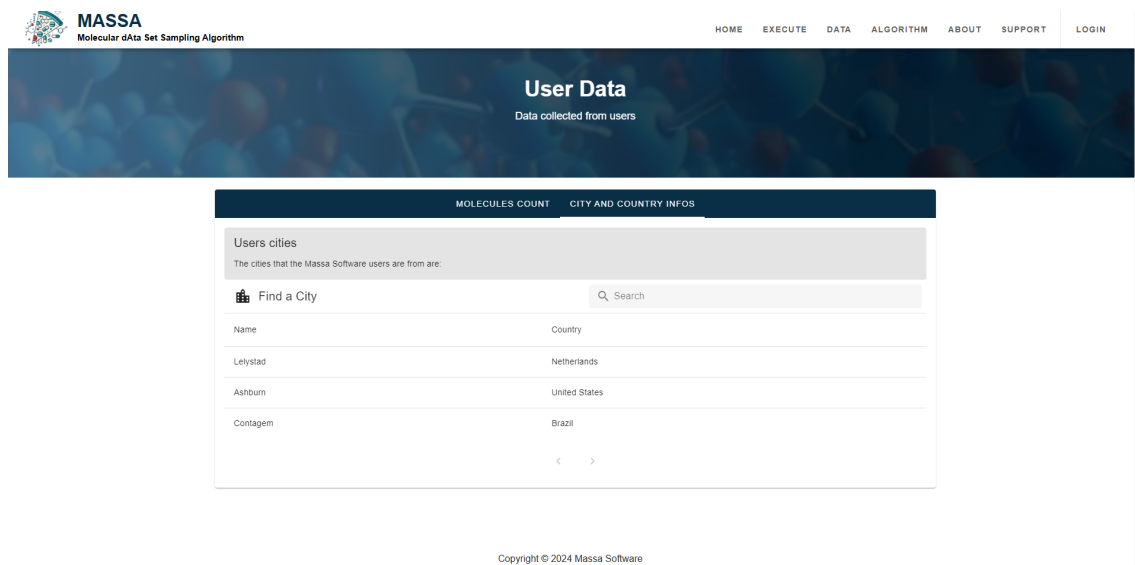
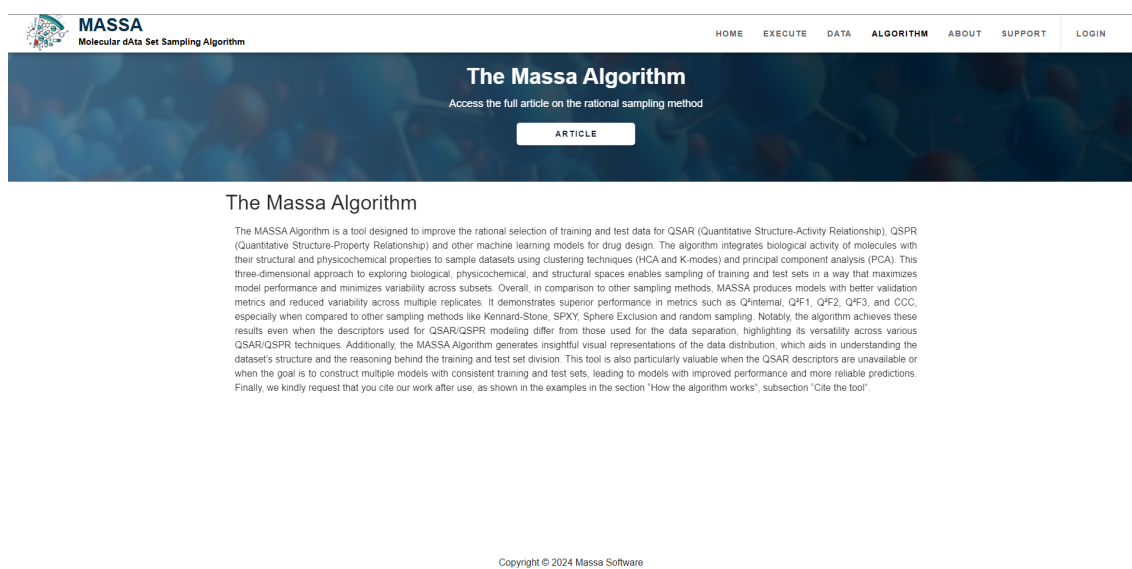


Figura 11. Application Users Location Page

A Figura 12 apresenta a tela de Algoritmo descreve brevemente o funcionamento do Algoritmo MASSA, uma ferramenta desenvolvida para aprimorar a seleção racional de dados de treinamento e teste para modelos QSAR (Quantitative Structure-Activity Relationship) e QSPR (Quantitative Structure-Property Relationship), além de outros modelos de aprendizado de máquina para design de fármacos. O algoritmo integra a atividade

biológica das moléculas com suas propriedades estruturais e físico-químicas para amostrar conjuntos de dados usando técnicas de agrupamento (HCA e K-modes) e análise de componentes principais (PCA). Essa abordagem tridimensional explora os espaços biológicos, físico-químicos e estruturais, permitindo uma amostragem de conjuntos de treinamento e teste que maximiza o desempenho do modelo e minimiza a variabilidade entre as réplicas.

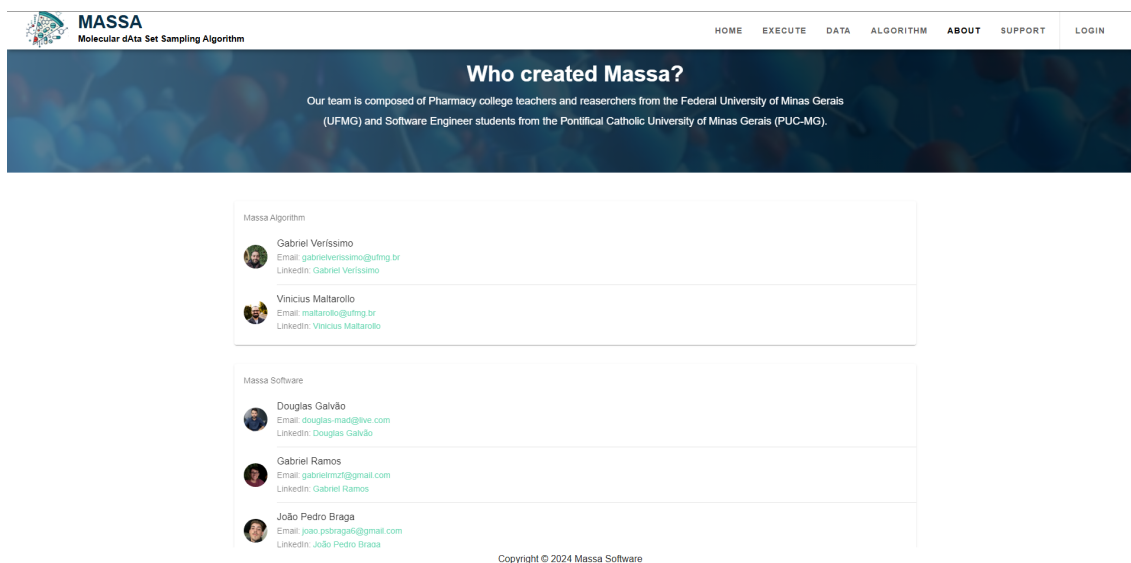
Um botão "ARTICLE" direciona o usuário para a publicação completa que descreve o algoritmo e seus resultados.



**Figura 12. Application Algorithm Page**

A Figura 13 apresenta a tela contendo as informações da equipe responsável pela criação do Software MASSA, dividindo os créditos entre o desenvolvimento do algoritmo e o desenvolvimento da aplicação web. A equipe do Algoritmo MASSA é composta por professores e pesquisadores da Faculdade de Farmácia da Universidade Federal de Minas Gerais (UFMG), enquanto a equipe do Software MASSA é formada por estudantes de Engenharia de Software da Pontifícia Universidade Católica de Minas Gerais (PUC-MG).

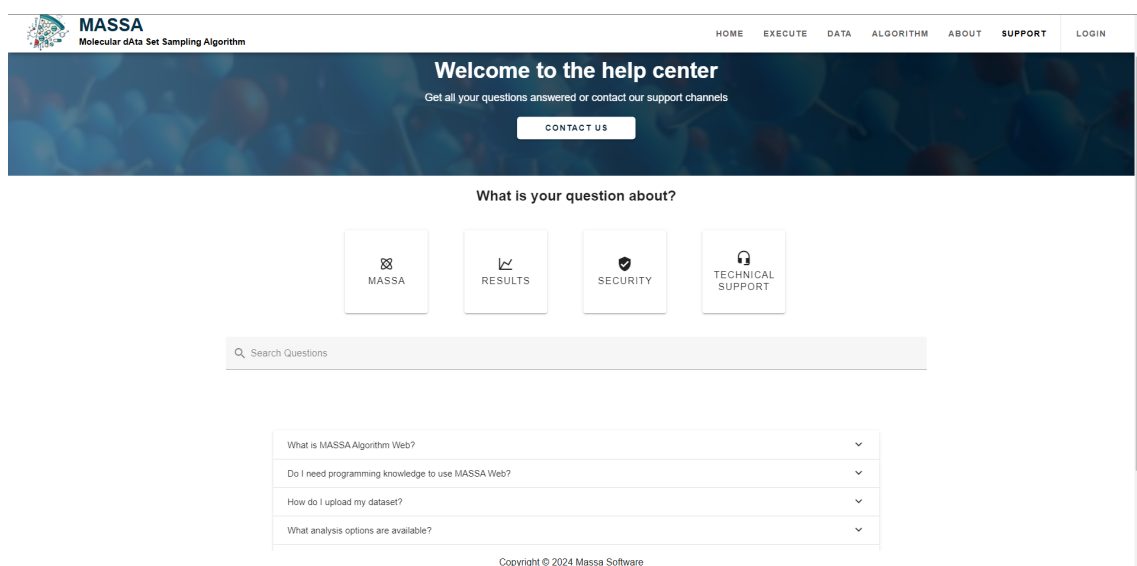
Para cada membro da equipe, são exibidos seu nome, foto, endereço de email e link para o perfil do LinkedIn. Essa apresentação clara e detalhada da equipe demonstra a colaboração interinstitucional entre UFMG e PUC-MG no desenvolvimento do projeto. A página mantém o padrão visual da aplicação. A estrutura da página facilita o contato com os responsáveis pelo projeto, promovendo transparência e incentivando a colaboração com a comunidade científica.



**Figura 13. Application About Page**

A página de Suporte, mostrada na Figura 14, apresenta a central de ajuda do software MASSA, oferecendo aos usuários recursos e opções de contato para suporte técnico e esclarecimento de dúvidas. A página é intitulada "Welcome to the help center" e convida o usuário a encontrar respostas para suas perguntas ou entrar em contato com os canais de suporte disponíveis. Um botão "CONTACT US" destaca a opção de contato direto com a equipe de suporte.

Abaixo, a seção "What is your question about?" apresenta quatro categorias principais de ajuda, representadas por ícones e texto: MASSA (informações gerais sobre o algoritmo), RESULTS (dúvidas sobre os resultados gerados), SECURITY (questões relacionadas à segurança dos dados) e TECHNICAL SUPPORT (suporte técnico para utilização da plataforma). Uma barra de pesquisa "Search Questions" permite que o usuário busque por palavras-chave relacionadas às suas dúvidas. Finalmente, uma lista de perguntas frequentes expansíveis oferece acesso rápido a informações relevantes sobre a plataforma, como "What is MASSA Algorithm Web?", "Do I need programming knowledge to use MASSA Web?", "How do I upload my dataset?" e "What analysis options are available?".



**Figura 14. Application Support Page**

## 5. Conclusões e trabalhos futuros

O presente trabalho teve como objetivo principal desenvolver um software para democratizar o acesso ao algoritmo MASSA, permitindo que usuários com qualquer nível de afinidade com o algoritmo, usem a ferramenta. Os objetivos específicos foram alcançados, incluindo a refatoração do código do algoritmo, a criação de uma interface web amigável, a implementação de um sistema robusto de login e autenticação, a geração de relatórios interativos que pudessem ser exportados em PDF, a integração de um tutorial na plataforma e a coleta de dados de uso para aprimoramento contínuo.

Os resultados demonstram que o desenvolvimento da aplicação web foi bem-sucedido, oferecendo uma solução acessível e intuitiva, que permita o uso do algoritmo MASSA de forma eficiente. A arquitetura do software, baseada em camadas, utilizando FastAPI e Vue.js, provou-se robusta e escalável, permitindo que a construção do software diante do surgimento e modelagem contínua de novos requisitos fosse possível. O sistema de login, com autenticação segura via JWT, protege as informações dos usuários. A capacidade de gerar relatórios interativos enriqueceram significativamente o processo de análise dos resultados do algoritmo. Além disso, a inclusão do tutorial e a coleta de dados de uso anônimo garantem maior usabilidade e direcionam futuras melhorias na interface e funcionalidades.

A aplicação, disponibilizada aqui, juntamente com seu repositório no GitHub, atende às necessidades da comunidade científica e do Laboratório de Farmácia da UFMG, permitindo uma maior inclusão em pesquisas QSAR/QSPR.

Além disso, alguns pontos ainda podem ser aprimorados em trabalhos futuros. A melhoria do sistema de fila de execução atual pode expandir as capacidades do software, pensando-se até na disponibilização de uma API RESTful para integração com outras ferramentas e plataformas computacionais.

Por fim, o projeto do MASSA demonstra a importância da interação entre a engenharia de software e a sociedade, onde por meio da tecnologia e das metodologias de desenvolvimento, pode-se proporcionar soluções que gerem valor de verdade. Nesse caso em específico, o software garante a disseminação e popularização de métodos computacionais avançados para a amostragem de dados em QSAR/QSPR.

O feedback do cliente, obtido por meio de um questionário, apresentou nota máxima em todas as questões, que foram relacionadas ao desempenho e comprometimento do time de desenvolvimento e sobre o resultado final do software.

**Tabela 3. Resumo das Respostas do Questionário**

Questão	Resposta
1.1 (Software atende necessidades)	5
1.2 (Resultados de acordo com expectativas)	5
1.3 (Recomendaria outra instituição para fazer esse projeto)	0
2.1 (Diálogo com os alunos)	5
2.2 (Interesse e envolvimento dos alunos)	5
2.3 (Aplicação de competências)	5
2.4 (Inovação na solução)	5

## Referências

- [Ferreira et al. 2022] Ferreira, R. d. A., Teixeira, G., and Peternelli, L. A. (2022). Kennard-stone method outperforms the random sampling in the selection of calibration samples in snps and NIR data. *Ciência Rural*, 52(5).
- [Martin 2009] Martin, R. C. (2009). *Clean Code: A Handbook of Agile Software Craftsmanship*. Prentice Hall, Upper Saddle River, NJ.
- [Todeschini and Consonni 2009] Todeschini, R. and Consonni, V. (2009). *QSAR Models for Prediction of Biological Activity: Methods and Techniques*. Wiley-VCH, Weinheim.
- [Van Rossum et al. 2001] Van Rossum, G., Warsaw, B., and Coghlan, A. (2001). PEP 8 – style guide for python code. Process 8, Python Software Foundation. Status: Active, Created: 05-Jul-2001, Post-History: 05-Jul-2001, 01-Aug-2013.
- [Veríssimo et al. 2023] Veríssimo, G. C., Pantaleão, S. Q., de Oliveira Fernandes, P., Gertrudes, J. C., Kronenberger, T., Honório, K. M., and Maltarollo, V. G. (2023). Massa algorithm: an automated rational sampling of training and test subsets for qsar modeling. In Maltarollo, V. G. and Veríssimo, G. C., editors, *Journal of Computer-Aided Molecular Design*. Springer.