

SAMPLE: DATASET FOCADO EM APRENDIZADO DE MÁQUINA PARA PESQUISA DE PEPTÍDEOS ANTIMICROBIANOS

LUIZA PEREZ CASAGRANDE¹; DANIELA PERES MARTINEZ²; DÉBORAH TROTA FARIAS DE ALBERNAZ³; FREDERICO SCHMITT KREMER⁴

¹Universidade Federal de Pelotas – lperezcasagrande@gmail.com

²Universidade Federal de Pelotas – daanmartinez2@gmail.com

³Universidade Federal de Pelotas – trotadeborah@gmail.com

⁴Universidade Federal de Pelotas – fred.s.kremer@gmail.com

1. DESCRIÇÃO DA INOVAÇÃO

A aceleração da ameaça global representada pela resistência antimicrobiana compromete crescentemente a eficácia dos antibióticos convencionais, tornando necessária a descoberta de novos peptídeos antimicrobianos (AMPs). Para enfrentar esse desafio, foi desenvolvido o sAMPlE, uma base de dados e plataforma integrada no formato ML-ready (pronta para aprendizado de máquina), projetada para organizar, padronizar e impulsionar a identificação e o desenvolvimento de novas moléculas no combate às infecções resistentes.

Em termos de funcionamento, o sAMPlE atua como um ecossistema que resolve a histórica fragmentação das informações sobre peptídeos, reunindo dados dispersos em artigos científicos, bancos de dados acadêmicos e depósitos de patentes industriais. Por meio de um fluxo automatizado alimentado por modelos de inteligência artificial multimodal, a ferramenta lê e extrai informações de documentos não estruturados, transformando-os em registros padronizados. Cada peptídeo cadastrado é automaticamente enriquecido com suas propriedades físico-químicas, estimativas de toxicidade, modelos de estrutura tridimensional e representações numéricas pré-calculadas. Toda essa estrutura é disponibilizada em uma plataforma web amigável e consultável, permitindo que pesquisadores busquem sequências por similaridade e executem modelos preditivos de forma simples e intuitiva, sem a necessidade de etapas complexas e demoradas de pré-processamento manual de dados.

Como diferenciais, o sAMPlE destaca-se por incluir de forma inédita o acervo de patentes industriais, preenchendo o chamado “patent gap” do setor ao incorporar sequências sintéticas e otimizadas que possuem maior estabilidade e menor toxicidade, que são aspectos raramente encontrados em bases acadêmicas tradicionais. Além disso, ao alinhar-se aos princípios FAIR (Localizável, Acessível, Interoperável e Reutilizável) e disponibilizar vetores de dados prontos para uso em modelos de inteligência artificial, a plataforma elimina gargalos computacionais e oferece uma infraestrutura completa e aberta para acelerar a descoberta de novos candidatos a fármacos.

2. ANÁLISE DE MERCADO

A inovação destina-se a pesquisadores acadêmicos, bioinformatas, programadores de modelos de *machine learning* (ML) e indústrias farmacêuticas focadas no desenvolvimento de novos peptídeos antimicrobianos (AMPs). Atualmente, o setor farmacêutico enfrenta o enorme desafio de reduzir o tempo e os elevados custos associados à descoberta e ao desenvolvimento de novos

medicamentos (Precedence Research, 2026). Como reflexo desta necessidade, mais de 68% das organizações farmacêuticas passaram a subcontratar fluxos de trabalho de bioinformática, e aproximadamente 72% dos programas de pesquisa farmacêutica dependem diretamente da análise de dados genômicos e proteômicos (Business Research Insights, 2026). A plataforma sAMPlE atende a este público ao oferecer um ecossistema preparado para inteligência artificial, com dados padronizados, modelos estruturais 3D e *embeddings* pré-calculados de alta dimensão, o que permite o treino imediato de redes preditivas sem a necessidade de elevado processamento inicial por parte do utilizador final.

O ecossistema de dados para AMPs encontra-se atualmente fragmentado em dois principais domínios concorrentes. Por um lado, existem os repositórios académicos, como o APD3, o CAMP e o DBAASP, que, embora apresentem uma elevada curadoria, costumam ter ciclos de atualização lentos, focam-se maioritariamente em sequências naturais e não são otimizados para algoritmos de *machine learning*, carecendo de anotações químicas padronizadas para resíduos modificados e de rigorosa deduplicação de dados. Por outro lado, os bancos de patentes industriais, como o Espacenet, abrangem um vasto volume de sequências sintéticas otimizadas, mas estas informações encontram-se ocultas em formatos não estruturados (como PDFs e XMLs), o que dificulta a extração sistemática e a interoperabilidade. A solução sAMPlE destaca-se da concorrência ao utilizar modelos de inteligência artificial multimodais (Gemini 2.5 Flash e Pro) para extrair, verificar e integrar ambos os domínios numa única infraestrutura unificada e adequada aos princípios FAIR (*Findable, Accessible, Interoperable, and Reusable*).

A resistência antimicrobiana (AMR) é uma ameaça fundamental à saúde pública, estando previsto que cause 10 milhões de mortes anuais até 2050 caso não sejam desenvolvidas novas soluções terapêuticas. Esta urgência tem impulsionado o financiamento da biologia computacional e a adoção de IA na indústria, definindo o mercado total endereçável (TAM) através do mercado global de Inteligência Artificial na Descoberta de Fármacos, que deverá crescer de 2,9 mil milhões de dólares em 2026 para 13,8 mil milhões de dólares até 2033, com uma taxa de crescimento anual composta (CAGR) de 24,8% (Grand View Research, 2026). O mercado disponível para serviço (SAM) enquadra-se no mercado global de Serviços de Bioinformática, avaliado em 4,76 mil milhões de dólares em 2026 e com previsão de atingir 19,14 mil milhões de dólares até 2035 (CAGR de 16,71%), sendo que o segmento específico de serviços para a descoberta de medicamentos representa aproximadamente 17% deste total (Business Research Insights, 2026). Por fim, o mercado passível de ser obtido (SOM) insere-se no mercado de vigilância da resistência antimicrobiana, que deverá crescer de 7,13 mil milhões de dólares em 2026 para 12,18 mil milhões de dólares em 2034 (CAGR de 6,92%) (Fortune Business Insights, 2026). O projeto sAMPlE visa capturar a parcela deste mercado representada por instituições de pesquisa e empresas de biotecnologia que procuram ativamente plataformas integradas e algoritmos preditivos baseados no modelo ESM-2.

3. ESTRATÉGIA DE DESENVOLVIMENTO E IMPLEMENTAÇÃO

O sAMPlE é uma infraestrutura científica aberta e gratuita financiada pelo Lacuna Fund. A plataforma e os seus artefactos (HDF5, ChromaDB, modelos 3D) estão disponíveis publicamente na web e no Zenodo. O valor gerado reside em

capacitar investigadores e empresas a desenvolverem as suas próprias inovações biotecnológicas, fármacos e patentes a partir destes dados abertos.

Todo o ecossistema (código, base de dados e ferramentas) possui licença aberta, seguindo os princípios FAIR. Ao agregar dados de literatura e patentes públicas, o sAMPlE garante total liberdade de operação. Inovações, moléculas otimizadas ou modelos de IA desenvolvidos por terceiros a partir do sAMPlE pertencem integralmente aos seus criadores, que mantêm pleno direito de patente e comercialização.

Atualmente, a infraestrutura possui um protótipo funcional demonstrado em ambiente operacional. O projeto já consolidou 133.916 fontes, foi desenvolvido o pipeline de extração por IA multimodal em dois estágios, realizada a validação técnica (padrão-ouro de 500 documentos) e lançada a plataforma web. Os próximos passos incluem atualizações contínuas da base e testes *in vitro* para a validação biológica de sequências promissoras.

Além disso, o risco de inconsistências na extração de dados é mitigado pela arquitetura adversarial da IA (Gemini 2.5 Pro a auditar o Gemini 2.5 Flash), validações rigorosas em JSON, verificações quimioinformáticas no RDKit e revisão humana. O risco de sustentabilidade computacional e custos é neutralizado pelo uso de uma arquitetura leve (base de dados SQLite num único ficheiro e HDF5 para vetores), assegurando a durabilidade da plataforma a longo prazo.

4. RESULTADOS ESPERADOS E IMPACTO

O impacto social e científico da inovação reside na democratização de uma base unificada de dados de peptídeos antimicrobianos, acelerando a identificação de moléculas terapêuticas contra infeções resistentes. No âmbito económico e operacional, o uso de plataformas de bioinformática e IA integradas reduz em cerca de 55% a 60% o tempo de processamento e análise de dados biológicos, eliminando custos redundantes de pré-processamento manual para laboratórios e empresas de P&D. Como visão de futuro ancorada no desenvolvimento da ferramenta, o sAMPlE projeta a automação contínua da ingestão de documentos de patentes e literatura em tempo real, a expansão de modelos preditivos e gerativos para o design *de novo* de peptídeos e o mapeamento de efeitos sinérgicos entre peptídeos e antibióticos convencionais extraídos da literatura

5. CONCLUSÕES

A plataforma sAMPlE representa um marco transformador na investigação de peptídeos antimicrobianos (AMPs), superando a histórica fragmentação de dados entre os repositórios académicos e as patentes industriais. Ao utilizar uma arquitetura de inteligência artificial multimodal inovadora para extrair e validar estruturalmente informações de um vasto corpus documental, a inovação estabelece um ecossistema "ML-ready" e estritamente compatível com os princípios FAIR. O fornecimento de dados altamente curados, juntamente com embeddings pré-calculados (ESM-2) e modelos 3D, elimina os principais gargalos computacionais da área. O impacto potencial desta infraestrutura é amplo, permitindo acelerar substancialmente a descoberta e o desenho preditivo de novos antibióticos de próxima geração frente à crescente ameaça global da resistência antimicrobiana (AMR).

Por fim, convidamos pesquisadores acadêmicos, startups de biotecnologia e empresas farmacêuticas a explorarem e utilizarem livremente esta infraestrutura pública de dados, para impulsionar os seus próprios fluxos de trabalho de machine learning. O envolvimento ativo da comunidade científica, tecnológica e de potenciais parceiros de financiamento é o próximo passo fundamental para traduzir este avanço da biologia computacional em soluções clínicas concretas e comercializáveis.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BUSINESS RESEARCH INSIGHTS. Visão Geral do Mercado de Serviços de Bioinformática. Business Research Insights, 2026. Acessado em 18 set. 2026. Online. Disponível em:
<https://www.businessresearchinsights.com/pt/market-reports/bioinformatics-services-market-129520>

FORTUNE BUSINESS INSIGHTS. Antimicrobial Resistance Surveillance Market Size and Industry Overview. Fortune Business Insights, 2026. Acessado em 18 set. 2026. Online. Disponível em:
<https://www.fortunebusinessinsights.com/antimicrobial-resistance-surveillance-market-112167>

GRAND VIEW RESEARCH. Artificial Intelligence In Drug Discovery Market Summary. Grand View Research, 2026. Acessado em 18 set. 2026. Online. Disponível em:
<https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-drug-discovery-market>

PRECEDENCE RESEARCH. What is the AI in Pharmaceutical Market Size?. Precedence Research, 2026. Acessado em 18 set. 2026. Online. Disponível em:
<https://www.precedenceresearch.com/ai-in-pharmaceutical-market>