UNIVERSIDADE FEDERAL DO PARANÁ

DIOGO DE JESUS SOARES MACHADO

INCORPORAÇÃO DE TEXTO BASEADA EM PROJEÇÃO ALEATÓRIA INSPIRADA EM BIOINFORMÁTICA

CURITIBA

2026

DIOGO DE JESUS SOARES MACHADO

INCORPORAÇÃO DE TEXTO BASEADA EM PROJEÇÃO ALEATÓRIA INSPIRADA EM BIOINFORMÁTICA

Tese apresentada ao Programa de Pós-Graduação Associado em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Bioinformática.

Orientador: Prof. Dr. Roberto Tadeu Raittz
Coorientador: Prof. Dr. Fábio de Oliveira Pedrosa

CURITIBA

2026

Programa de Pós-Graduação
Associado em Bioinformática, Setor de Educação
Profissional e Tecnológica, Universidade Federal do Paraná

Catalogação na publicação
Sistema de Bibliotecas UFPR

M149

Machado, Diogo de Jesus Soares
Incorporação de texto baseada em projeção aleatória inspirada em bioinformática / Diogo de Jesus Soares Machado. - Curitiba, 2026.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós-Graduação Associado em Bioinformática UFPR/UTFPR-CP, 2026.

Orientador: Prof. Dr. Roberto Tadeu Raittz
Coorientador: Prof. Dr. Fábio de Oliveira Pedrosa

1. Bioinformática. 2. Processamento de linguagem natural. 3. Método de projeção aleatória. 4. Exploração de literatura. 5. Modelagem linguística. 6. Mineração de textos. I. Raittz, Roberto Tadeu. II. Pedrosa, Fábio de Oliveira. III. Título. IV. Universidade Federal do Paraná V. Universidade Tecnológica Federal do Paraná.

CDD 570.285

Bibliotecário: André A. Oliveira — CRB 9/2019

ATA Nº90

ATA DE SESSÃO PÚBLICA DE DEFESA DE DOUTORADO PARA A OBTENÇÃO DO GRAU DE DOUTOR EM BIOINFORMÁTICA

No dia tres de novembro de dois mil e vinte e cinco às 09:00 horas, na sala Auditório - Bloco A, Setor de Educação Profissional e Tecnológica - UFPR - Rua Doutor Alcides Vieira Arcoverde, 1225 Jardim das Américas - Curitiba-PR-BR - CEP: 81520-260, foram instaladas as atividades pertinentes ao rito de defesa de tese do doutorando **DIOGO DE JESUS SOARES MACHADO**, intitulada: **INCORPORAÇÃO DE TEXTO BASEADA EM PROJEÇÃO ALEATÓRIA INSPIRADA EM BIOINFORMÁTICA**. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação BIOINFORMÁTICA da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: ALEXANDER ROBERT KUTZKE (UNIVERSIDADE FEDERAL DO PARANÁ), ROBERTO HIROCHI HERAI (PONTIFICA UNIVERSIDADE CATÓLICA DO PARANA), DIEVAL GUIZELINI (UNIVERSIDADE FEDERAL DO PARANÁ), EDUARDO TIEPPO (INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PARANÁ - IFPR). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela APROVAÇÃO. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de doutor está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, ALEXANDER ROBERT KUTZKE, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

Curitiba, 03 de Novembro de 2025.

Assinatura Eletrônica
05/02/2026 14:34:06.0
ALEXANDER ROBERT KUTZKE
Presidente da Banca Examinadora

Assinatura Eletrônica
05/02/2026 12:05:52.0
ROBERTO HIROCHI HERAI
Avaliador Externo (PONTIFICA UNIVERSIDADE CATÓLICA DO PARANA)

Assinatura Eletrônica
04/02/2026 15:39:01.0
DIEVAL GUIZELINI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
04/02/2026 15:39:48.0
EDUARDO TIEPPO
Avaliador Externo (INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PARANÁ - IFPR)

Assinatura Eletrônica
05/02/2026 11:12:01.0
FABIO DE OLIVEIRA PEDROSA
Coorientador(a) (UNIVERSIDADE FEDERAL DO PARANÁ)

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **DIOGO DE JESUS SOARES MACHADO**, intitulada: **INCORPORAÇÃO DE TEXTO BASEADA EM PROJEÇÃO ALEATÓRIA INSPIRADA EM BIOINFORMÁTICA**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 03 de Novembro de 2025.

Assinatura Eletrônica
05/02/2026 14:34:06.0
ALEXANDER ROBERT KUTZKE
Presidente da Banca Examinadora

Assinatura Eletrônica
05/02/2026 12:05:52.0
ROBERTO HIROCHI HERAI
Avaliador Externo (PONTIFICA UNIVERSIDADE CATÓLICA DO PARANA)

Assinatura Eletrônica
04/02/2026 15:39:01.0
DIEVAL GUIZELINI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
04/02/2026 15:39:48.0
EDUARDO TIEPPO
Avaliador Externo (INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PARANÁ - IFPR)

Assinatura Eletrônica
05/02/2026 11:12:01.0
FABIO DE OLIVEIRA PEDROSA
Coorientador(a) (UNIVERSIDADE FEDERAL DO PARANÁ)

A todo coletivo.

AGRADECIMENTOS

Agradeço àquela que é tudo e todos, conhecida por muitos nomes, mas que aqui invoco como Natureza.

“A beleza de uma tecnologia ecológica – uma ecotecnologia, ou uma tecnologia libertária, ou uma tecnologia alternativa – é que as pessoas podem entendê-la se estiverem dispostas a tentar dedicar algum grau de esforço para isso. É a simplicidade, sempre que possível, é a pequena escala, sempre que possível. É disso que estou falando. Não estou falando em voltar ao paleolítico, não estou falando em voltar para as cavernas. Não podemos voltar a isso e acho que não queremos voltar a isso.”

(Murray Bookchin)

RESUMO

Este estudo apresenta o SWeePtex, uma metodologia para a geração de incorporações vetoriais de texto (*text embeddings*) por meio de projeção aleatória, inspirada em técnicas de Bioinformática. A abordagem permite a criação de modelos a partir do zero (*from scratch*), sendo particularmente útil em domínios específicos. O SWeePtex adapta o método SWeeP (*Spaced Words Projection*), originalmente concebido para sequências biológicas, partindo da premissa de que a linguagem natural e as sequências biológicas compartilham uma estrutura comum de sequências de entidades elementares. Esta analogia é formalizada por meio do conceito de texto como uma sequência biológica (*Biological Sequence-Like*, BSL), na qual textos são codificados no formato FASTA para a aplicação direta de métodos de Bioinformática. A proposta é desenvolvida por meio de três artigos: o Artigo 1 introduz o *framework* Biotext, que integra o SWeePtex por meio da manipulação de textos em BSL; o Artigo 2 apresenta uma avaliação quantitativa por meio de uma plataforma de comparação; e o Artigo 3 apresenta o TXTree (*Text Tree*), um gerador de interface portátil para a exploração de literatura. Consequentemente, o SWeePtex contribui para uma perspectiva epistemológica alternativa na modelagem de linguagem, fundamentada em princípios matemáticos e de representação distintos dos paradigmas de aprendizado profundo predominantes. Como resultado, o SWeePtex estabelece-se como uma alternativa viável, atuando como um contraponto construtivo e um catalisador de soluções futuras. Qualitativamente, sua viabilidade e relevância são atestadas por meio de exemplos de uso e de uma publicação científica revisada por pares. Quantitativamente, embora resultados preliminares o mostrem comparável a modelos neurais compactos, reconhece-se que barreiras metodológicas de avaliação permanecem e devem ser abordadas em projetos futuros. Assim, o SWeePtex demonstra a generalização bem-sucedida do SWeeP para além do seu domínio original, posicionando-o como uma técnica baseada no paradigma da projeção aleatória com potencial abrangente. Para fomentar avanços, o *software* está publicamente disponível em duas implementações: o pacote Biotext no PyPI (`https://pypi.org/p/biotext`) e a aplicação TXTree no SourceForge (`https://sf.net/p/txtree`).

Palavras-chaves: Bioinformática. Processamento de linguagem natural. Método de projeção aleatória. Exploração de literatura. Modelagem linguística. Mineração de textos.

ABSTRACT

This study presents SWeePtex, a methodology for generating text embeddings via random projection, inspired by Bioinformatics techniques. The approach enables the creation of models from scratch, proving particularly useful for specific domains. SWeePtex adapts the SWeeP method (Spaced Words Projection), originally conceived for biological sequences, based on the premise that natural language and biological sequences share a common structure of elementary entity sequences. This analogy is formalized through the concept of text as a biological sequence (Biological Sequence-Like, BSL), where texts are encoded in the FASTA format for the direct application of Bioinformatics methods. The proposal is developed across three articles: Article 1 introduces the Biotext framework, which integrates SWeePtex by manipulating BSL texts; Article 2 provides a quantitative evaluation through a benchmarking platform; and Article 3 presents TXTree (Text Tree), a portable interface generator for literature exploration. Consequently, SWeePtex contributes to an alternative epistemological perspective in language modeling, grounded in mathematical and representational principles distinct from prevailing deep learning paradigms. As a result, SWeePtex establishes itself as a viable alternative, serving as a constructive counterpoint and a catalyst for future solutions. Qualitatively, its feasibility and relevance are supported by usage examples and a peer-reviewed scientific publication. Quantitatively, although preliminary results indicate it is comparable to compact neural models, methodological evaluation barriers remain and must be addressed in future projects. Thus, SWeePtex demonstrates the successful generalization of the SWeeP beyond its original domain, positioning it as a random-projection-based technique with broad potential. To foster progress, the software is publicly available in two implementations: the Biotext package on PyPI (`https://pypi.org/p/biotext`) and the TXTree application on SourceForge (`https://sf.net/p/txtree`).

Keywords: Bioinformatics. Natural language processing. Random projection method. Literature exploration. Linguistic modeling. Text mining.

LISTA DE ABREVIATURAS E DE SIGLAS

AI	*Artificial Intelligence* (Inteligência Artificial)
AMINOcode	*Amino Acid Code* (Código de Aminoácidos)
ASCII	*American Standard Code for Information Interchange* (Código Padrão Americano para o Intercâmbio de Informação)
BERT	*Bidirectional Encoder Representations from Transformers* (Representações de Codificador Bidirecional de Transformadores)
BSL	*Biological Sequence-Like* (Semelhante a Sequência Biológica)
CBOW	*Continuous Bag of Words* (Saco de Palavras Contínuo)
CPU	*Central Processing Unit* (Unidade Central de Processamento)
DNA	*Deoxyribonucleic Acid* (Ácido Desoxirribonucleico)
DNAbits	*Deoxyribonucleic Acid Bits* (Bits de Ácido Desoxirribonucleico)
DeepSeek-R1	*DeepSeek Reasoner 1* (Raciocinador DeepSeek 1)
GPT	*Generative Pre-trained Transformer* (Transformador Pré-treinado Generativo)
HDF5	*Hierarchical Data Format version 5* (Formato de Dados Hierárquico versão 5)
HDV	*High-Dimensional Vector* (Vetor de Alta Dimensionalidade)
HTML	*HyperText Markup Language* (Linguagem de Marcação de Hipertexto)
HTML-TM	*HyperText Markup Language for Text Mining* (Linguagem de Marcação de Hipertexto para Mineração de Textos)
IDF	*Inverse Document Frequency* (Frequência Inversa do Documento)
LDV	*Low-Dimensional Vector* (Vetor de Baixa Dimensionalidade)
LLM	*Large Language Model* (Modelo de Linguagem de Grande Escala)

MEDLINE	*Medical Literature Analysis and Retrieval System Online* (Sistema de Análise e Recuperação de Literatura Médica Online)
MLM	*Masked Language Modeling* (Modelagem de Linguagem Mascarada)
MLP	*Multi Layer Perceptron* (Perceptron Multicamadas)
MTEB	*Massive Text Embedding Benchmark* (Referência para Avaliação de Incorporação de Textos em Larga Escala)
NCBI	*National Center for Biotechnology Information* (Centro Nacional de Informação Biotecnológica)
NLM	*Natural Language Modeling* (Modelagem de Linguagem Natural)
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
PCA	*Principal Component Analysis* (Análise de Componentes Principais)
PMID	*PubMed Identifier* (Identificador do PubMed)
PyPI	*Python Package Index* (Índice de Pacotes do Python)
RAM	*Random Access Memory* (Memória de Acesso Aleatório)
RNN	*Recurrent Neural Network* (Rede Neural Recorrente)
SWeeP	*Spaced Words Projection* (Projeção de Palavras Espaçadas)
SWeePtex	*SWeeP for Texts* (SWeeP para Textos)
SWeePtex-Emb	*SWeeP Text Embedding* (Incorporação de Textos com SWeeP)
Skip-gram	*Continuous Skip-gram* (Skip-gram Contínuo)
T5	*Text-to-Text Transfer Transformer* (Transformador de Transferência Texto-para-Texto)
TF	*Term Frequency* (Frequência do Termo)
TF-IDF	*Term Frequency – Inverse Document Frequency* (Frequência do Termo – Frequência Inversa do Documento)

TM	*Text Mining* (Mineração de Textos)
TXTree	*Text Tree* (Árvore de Textos)
WEBSOM	*Web Self-Organizing Map* (Mapa Auto-Organizável da Web)
Word2Vec	*Word to Vector* (Palavra para Vetor)
nDCG@10	*Normalized Discounted Cumulative Gain at 10* (Ganho Cumulativo Descontado Normalizado em 10)

SUMÁRIO

1 INTRODUÇÃO

Os Grandes Modelos de Linguagem (*Large Language Models*, LLMs) representam um paradigma vigente no processamento de linguagem natural e na geração de representações vetoriais de texto (*embeddings*). Sua aplicação prática, contudo, enfrenta desafios substanciais. A abordagem de aprendizado profundo (*deep learning*), além de exigir custos computacionais proibitivos em muitos cenários (Schwartz et al., 2020), é inerentemente opaca. Essa falta de transparência constitui um obstáculo epistemológico: sem compreender como o significado é construído a partir dos textos, torna-se impossível realizar otimizações estruturais profundas ou conceber alternativas eficientes. O problema é agravado pela tendência de modelos excessivamente generalistas de produzir representações superficiais ou enviesadas (Currie, 2023), o que reforça a urgência por métodos com funcionamento interno transparente e auditável (Scorzato, 2024).

No campo da pesquisa científica, há a tendência de executar LLMs localmente, buscando maior privacidade, custos reduzidos e reprodutibilidade (Hutson, 2024). Ferramentas como o Ollama (Ollama, 2026) viabilizam a execução local de modelos como Llama, Phi e Gemma. Contudo, essa prática apenas mitiga questões de infraestrutura, sem abordar o cerne do problema: a criação e o refinamento (*fine-tuning*) de modelos nesse paradigma permanecem processos extensivos e computacionalmente custosos (Li et al., 2025; Yao et al., 2022). A barreira persiste porque a complexidade e a opacidade do mecanismo de aprendizado profundo dificultam uma reengenharia eficiente. Portanto, a exploração de paradigmas alternativos torna-se conveniente e necessária. Um caminho promissor é o desenvolvimento de métodos cujo mecanismo de formação de vetores seja matematicamente simples e transparente, permitindo não apenas uma execução eficiente, mas também uma compreensão teórica que guie a criação de modelos sustentáveis em domínios específicos.

Como paradigma alternativo para a incorporação vetorial de textos, a projeção aleatória possibilita a interpretabilidade por meio de operações matemáticas diretas e transparentes. Seu fundamento teórico é o lema de Johnson-Lindenstrauss (Johnson; Lindenstrauss, 1984), que garante a preservação aproximada das distâncias entre vetores após sua projeção em um espaço de dimensão inferior, ainda suficientemente grande para assegurar a representatividade dos dados. Neste processo, a projeção aleatória atua como um mecanismo de extração de características (*feature extraction*): as combinações definidas pela matriz de projeção criam novas dimensões sintéticas que capturam padrões e relações presentes na representação inicial de alta dimensionalidade do texto. Dessa forma, o método funciona como um processo auditável, transformando a alta dimensionalidade dos textos em representações compactas e es-

truturalmente consistentes, nas quais cada etapa da transformação é explicitamente compreensível (Ganguli; Sompolinsky, 2012).

Analogamente à cognição animal, a projeção aleatória funciona como um mecanismo de abstração. Ela combina aleatoriamente características (*features*) sensoriais de alta dimensionalidade para formar representações mentais compactas (*embeddings*). Embora envolva uma perda seletiva de informação, essa compressão preserva relações fundamentais de similaridade estrutural. Isso remete à hipótese de codificação eficiente de Barlow (Barlow; Rosenblith, 1961). Conforme esta hipótese, tanto sistemas biológicos quanto artificiais podem comprimir dados sensoriais em representações de menor dimensionalidade. Este processo filtra ruído e retém informações semanticamente relevantes, assegurando, assim, a eficiência computacional (Ganguli; Sompolinsky, 2012).

O método SWeeP (*Spaced Words Projection*) (De Pierri et al., 2020) aplica projeções aleatórias à representação de sequências biológicas, com sua eficácia consolidada em diversos estudos (Silva Filho et al., 2021; De Pierri et al., 2020; Perico et al., 2022; Raittz et al., 2021). Dada uma adaptação adequada, sua aplicação pode ser estendida ao domínio da linguagem natural. Na implementação original, o SWeeP processa entradas no formato FASTA, que utiliza cadeias de caracteres para representar dados biológicos. Portanto, ao codificar textos em formato análogo ao de sequências biológicas (*Biological Sequence-Like*, BSL), viabiliza-se a criação do SWeeP para textos (SWeePtex). Esta transposição conceitual da Bioinformática permite incorporar características epistemológicas frequentemente desejáveis na área, estabelecendo, assim, um novo caminho para a investigação.

Este estudo apresenta três artigos que detalham a concepção, aplicação e avaliação do SWeePtex, com a finalidade de analisar o comportamento em relação às problemáticas mencionadas e a contribuição para o paradigma da projeção aleatória. O primeiro[1] introduz a metodologia, descreve o pacote de programação e fornece um exemplo de uso. O segundo[2] apresenta uma análise quantitativa que compara a abordagem com modelos baseados em redes neurais artificiais. Por fim, o terceiro[3] demonstra a aplicação prática por meio do TXTree, uma ferramenta em linha de co-

[1]Artigo 1 – *Biotext with SWeePtex: Bioinformatics Tricks to Perform Fast, Accurate, and Content-specific String Embedding* (Biotext com SWeePtex: Truques de Bioinformática para Realizar Incorporação de Textos Rápida, Precisa e Específica ao Conteúdo)

[2]Artigo 2 – *SWeePtex-Emb: Benchmarking Random Projection-Based Text Embeddings Inspired by Bioinformatics* (SWeePtex-Emb: Avaliação Comparativa de Incorporações de Texto Baseadas em Projeção Aleatória Inspiradas pela Bioinformática)

[3]Artigo 3 – *TXTree: A Visual Tool for PubMed Literature Exploration by Text Mining* (TXTree: Uma Ferramenta Visual para Exploração de Literatura do PubMed por Mineração de Textos)

mando que viabiliza a geração de uma interface visual portátil para exploração de literatura baseada em HTML (*HyperText Markup Language*), denominada HTML-TM (*HTML for Text Mining*).

1.1 MOTIVAÇÃO

Reconhece-se a necessidade de desenvolver métodos para a incorporação de textos (*text embedding*) que priorizem a interpretabilidade e sejam, simultaneamente, computacionalmente acessíveis. O avanço dos Grandes Modelos de Linguagem (LLMs) baseados em redes neurais profundas trouxe capacidades notáveis, mas também consolidou um paradigma no qual o processo exato de formação dos vetores de incorporação é complexo e frequentemente não linear. Para favorecer a compreensão teórica e o desenvolvimento científico, é fundamental que se compreendam e se descrevam claramente os mecanismos pelos quais as representações são construídas, desde os dados de entrada até os vetores de saída.

A motivação primária deste trabalho é a busca por um paradigma alternativo de representação que promova a transparência metodológica e a compreensibilidade. Essa clareza teórica é um requisito para a criação de ferramentas mais acessíveis e reprodutíveis. Quando o processo de formação dos vetores é matematicamente explícito e cada etapa da transformação textual é logicamente compreensível, viabiliza-se a implementação de sistemas que podem ser auditados, replicados e adaptados com confiança pela comunidade de pesquisa. Esse entendimento detalhado permite derivar novos modelos, testar hipóteses linguísticas de forma controlada e explorar e refinar a relação entre a estrutura textual e a representação matemática.

A eficiência computacional constitui uma motivação consequente nesta proposta. Ela surge como um benefício característico de abordagens que, por serem fundamentadas em operações matemáticas diretas e bem definidas, dispensam processos de otimização iterativa de alto custo.

Diante do objetivo de favorecer uma compreensão teórica que viabilize ferramentas acessíveis e reprodutíveis, motiva-se a investigação de paradigmas com bases matemáticas explícitas e transparentes. A inspiração da Bioinformática é especialmente pertinente, pois este campo já oferece um conjunto de métodos fundamentados e interpretáveis para a representação de sequências de caracteres. Este legado fornece um caminho claro e reprodutível para modelar a linguagem natural de forma acessível ao escrutínio e ao desenvolvimento científico.

1.2 JUSTIFICATIVA

A adaptação de métodos da Bioinformática para o processamento de linguagem natural é uma abordagem metodologicamente sólida. Essa área já oferece técnicas comprovadas para a representação e a comparação de sequências de caracteres, cuja estrutura fundamental é análoga à dos textos. Tanto as sequências biológicas (de nucleotídeos ou aminoácidos) quanto as linguísticas (de letras e palavras) são definidas pela ordem de seus elementos constituintes, que determina sua função ou significado. Essa similaridade estrutural fundamenta a transposição de um domínio para o outro.

A abordagem central do SWeePtex é a projeção aleatória, que atua simultaneamente como mecanismo de extração de características e de redução de dimensionalidade. O fundamento teórico é garantido pelo lema de Johnson-Lindenstrauss, que assegura a preservação aproximada das relações estruturais entre os dados. Diferentemente do processo de treinamento de redes neurais, que é iterativo e gera representações indiretas, a projeção aleatória gera vetores por meio de operações aritméticas lineares aplicadas a uma matriz fixa. Este processo direto confere ao método uma transparência intrínseca, tornando explícita, auditável e matematicamente interpretável a transformação completa de texto em vetor.

A implementação desses princípios no método SWeePtex viabiliza, de forma crítica, uma compreensão lógica e completa da representação. Cada operação, desde a decomposição do texto em k-mers até a projeção final, mantém uma relação interpretável com os dados originais. Esta rastreabilidade é essencial para aplicações científicas que exigem validade analítica, reprodutibilidade e a capacidade de explicitar e testar os mecanismos subjacentes à representação linguística.

1.3 OBJETIVOS

1.3.1 Objetivo geral

Propor, desenvolver e validar a transposição do método SWeeP, baseado em projeção aleatória e originário da Bioinformática, para a criação de representações vetoriais de textos (SWeePtex). Esta validação, ancorada na premissa comum de tratar sequências de elementos interdependentes, materializa-se em três artigos científicos que constituem o núcleo desta tese.

1.3.2 Objetivos específicos

- Transpor e implementar o método: desenvolver e disponibilizar um pacote de *software* documentado que implemente a adaptação do algoritmo SWeeP para dados textuais (SWeePtex), garantindo usabilidade e reprodutibilidade. Este desenvolvimento constitui o núcleo do Artigo 1.
- Avaliar quantitativamente: realizar uma avaliação comparativa (*benchmark*) do desempenho do SWeePtex, contrastando sua efetividade e eficiência computacional com métodos do estado da arte baseados em redes neurais artificiais. Esta análise é apresentada no Artigo 2.
- Demonstrar a aplicabilidade prática: validar a utilidade do método em um cenário real por meio do desenvolvimento de uma ferramenta para a exploração de literatura científica (TXTree). Esta demonstração consolida a aplicabilidade do SWeePtex e é detalhada no Artigo 3.
- Discutir a contribuição: com base nos três artigos, delinear o panorama atual e o potencial do SWeePtex.

2 RESULTADO

O resultado deste estudo é composto por três artigos, escritos na língua inglesa, que atendem aos objetivos específicos delineados.

O Artigo 1 concretiza o primeiro objetivo específico ao propor formalmente a adaptação do algoritmo SWeePtex. Introduz o pacote de *software* implementado em Python, documentado e com exemplo de uso, que viabiliza a criação de incorporações (*embeddings*) a partir do zero (*from-scratch*), garantindo usabilidade, reprodutibilidade e interpretabilidade.

O Artigo 2 atende ao segundo objetivo específico ao apresentar uma avaliação comparativa (*benchmark*) com modelos baseados em redes neurais artificiais. A análise identifica desafios metodológicos no processo comparativo que precisam ser abordados em estudos futuros, mas também destaca potenciais quantitativos relevantes para o método, bem como caminhos a seguir.

O Artigo 3 cumpre o terceiro objetivo específico por meio do desenvolvimento da ferramenta TXTree, que aplica o SWeePtex para converter resultados de buscas do PubMed em uma interface visual interativa (HTML-TM). A ferramenta valida a utilidade do método em um cenário real de exploração de literatura, sendo operável localmente e fornecendo vetores para análises programáticas adicionais.

Cada um dos artigos, reproduzidos integralmente nas seções subsequentes, é precedido por uma breve contextualização em língua portuguesa. Estas contextualizações resumem o conteúdo e a contribuição.

Após a apresentação dos artigos, o quarto objetivo específico é atingido no capítulo de discussão, que integra as contribuições e os entendimentos alcançados com o propósito de explorar criticamente as implicações teóricas e práticas, analisando os potenciais e as limitações do método SWeePtex em sua situação atual.

2.1 ARTIGO 1 – BIOTEXT COM SWEEPTEX: FUNDAMENTAÇÃO METODOLÓGICA

O Artigo 1 introduz o *framework* Biotext, que adapta técnicas da Bioinformática ao domínio do Processamento de Linguagem Natural (NLP). Seu cerne metodológico reside na reconfiguração do algoritmo SWeeP, originalmente concebido para a análise de sequências biológicas, para a representação vetorial de textos. Essa transposição é viabilizada por duas estratégias de codificação distintas: o AMINOcode, que gera uma representação compacta por perda irreversível de informação, e o DNAbits, que preserva a codificação original sem perdas.

O método resultante, denominado SWeePtex, opera em duas etapas principais. Primeiro, constrói vetores de alta dimensionalidade a partir da contagem de padrões de palavras espaçadas na sequência codificada. Em seguida, aplica uma projeção pseudoaleatória conforme o lema de Johnson-Lindenstrauss. Esta etapa não é concebida meramente como redução de dimensionalidade, mas sim como uma técnica ativa para a captura e incorporação de características semanticamente relevantes.

Complementarmente, o *framework* inclui uma técnica de incorporação (*embedding*) contextual por média, aplicável a palavras e documentos. Inspirada no princípio distribucional, que postula que o significado de uma palavra é definido pelos contextos em que ocorre (Firth, 1957), esta abordagem permite que a representação lexical transcenda a ocorrência isolada, absorvendo o contexto semântico compartilhado entre os documentos em que a palavra aparece.

O artigo posiciona e conceitua a proposta com base em uma revisão da literatura sobre técnicas de modelagem de vetores linguísticos. Para contextualizá-la e evidenciar seus contrapontos, são apresentados tanto métodos recentes baseados em aprendizado profundo quanto exemplos do paradigma da aleatoriedade. Destes últimos, destaca-se o estudo de Kanerva (1994), que estabelece uma relação fundamental entre processos cognitivos biológicos e a vetorização mediada pela aleatoriedade.

Uma contribuição da abordagem reside na interpretabilidade do método. Diferentemente de modelos de aprendizagem profunda que funcionam como sistemas opacos, o SWeePtex mantém um vínculo direto e transparente entre características textuais originais e representações vetoriais finais, permitindo o endereçamento por conteúdo e maior controle analítico. Além disso, as etapas de geração dos vetores baseiam-se em operações conhecidas e teoricamente justificadas.

A formalização e a análise da complexidade computacional das abordagens do Biotext são apresentadas em nível teórico, incluindo AMINOcode, DNAbits, SWeeP-

tex e a incorporação com SWeePtex. A partir da teorização, justifica-se o aspecto do custo computacional, embora se destaque a importância de uma análise assintótica empírica como pesquisa futura.

O texto evidencia a publicação prévia sobre Yoga (Ieger-Raittz et al., 2025) como validação qualitativa da abordagem, mas também inclui uma nova implementação com o pacote Python e um exemplo de uso com 14.984 resumos da MEDLINE sobre tioredoxina. O resultado demonstra a capacidade do método de identificar agrupamentos semânticos de termos especializados, estruturar a literatura em grupos temáticos e reconhecer, de forma não supervisionada, vizinhanças semânticas, mantendo a rastreabilidade das representações.

O SWeePtex é especialmente destacado por permitir a modelagem do zero (*from-scratch*). Essa capacidade é crucial em domínios especializados em que modelos pré-treinados sobre o tema de interesse não estão disponíveis. A abordagem garante integridade referencial intrínseca ao processo, facilitando o rastreamento e a interpretabilidade dos resultados, características fundamentais para aplicações que demandam transparência analítica e controle sobre a representação final dos dados. Além disso, a interpretabilidade favorece a compreensão epistemológica da modelagem de vetores.

Biotext with SWeePtex: Bioinformatics Tricks to Perform Fast, Accurate, and Content-specific String Embedding

Diogo de J. S. Machado[1], Camilla R. De Pierri[1], Antonio C. da Silva Filho[1], Flávia de F. Costa[1], Nelson A. de M. Lemos[1], Camila P. Perico[1], Letícia G. C. Santos[1], Maricel G. Kann[2], Fábio de O. Pedrosa[1], and Roberto T. Raittz[1]

[1]Federal University of Paraná (UFPR)
[2]University of Maryland

Abstract

The growth of scientific literature underscores the need for efficient, interpretable methods for text analysis, particularly in specialized domains. While large language models (LLMs) offer broad capabilities, they often lack transparency and require substantial computational resources, making them less suitable for focused, domain-specific problems that require models to be built from scratch. This study presents Biotext, a novel framework that bridges Bioinformatics and Natural Language Processing (NLP) to address this gap. Built on three integrated pillars: theoretical design, software implementation, and experimental validation, Biotext introduces the SWeePtex method. SWeePtex first encodes text into a Biological Sequence-Like (BSL) format, then applies the proven Spaced Words Projection (SWeeP) algorithm with random projection, a technique originally developed for biological sequence analysis. This process generates high-dimensional, content-addressable vectors in near-linear time, eliminating the need for iterative model training. In an unsupervised usage example, analyzing 14,984 MEDLINE abstracts on thioredoxin, SWeePtex successfully organized biomedical terminology along a specificity gradient and clustered documents into semantically coherent research domains. The resulting embeddings, grounded in distributional principles via contextual averaging, enable intuitive semantic exploration and direct vector manipulation. By prioritizing transparency, scalability, and immediate applicability, SWeePtex offers a practical alternative for domain-specific text analysis from scratch. The Biotext Python package is freely available on PyPI (`https://pypi.org/p/biotext`).

Keywords: Text mining. Vector embedding. Bioinformatics. Random projection.

1 Introduction

Texts are the traditional way of storing scientific knowledge. However, the ever-growing volume of textual data in public databases underscores the urgent need for improved techniques for text manipulation and analysis (Tshitoyan et al. 2019). In this context, Text Mining (TM) methodologies are crucial, as they bridge human language and computational techniques to extract information and uncover hidden insights within complex texts (Hassani et al. 2020; Jurafsky and Martin 2025). This analytical process finds a significant methodological parallel in Bioinformatics. Both fields fundamentally process sequential symbolic data: TM analyzes strings of characters encoding linguistic information, while Bioinformatics analyzes strings of characters representing nucleotide or amino acid sequences, often in standardized formats such as FASTA.

This shared foundation in pattern recognition within symbolic sequences is not coincidental. Historically, Bioinformatics has integrated and refined techniques from diverse computational fields, including Computational Linguistics, to handle its specific data structures. This established cross-disciplinary adaptation sets a valuable precedent. Consequently, the sophisticated computational framework developed by Bioinformatics for the analysis, alignment, and interpretation of biological sequences presents a directly relevant and technically robust toolkit. Elements of this toolkit can be adapted and reapplied to address methodological challenges in TM, particularly in sequence modeling, feature extraction, and similarity detection within textual data.

Methods developed to map biological sequences to vector representations (Asgari and Mofrad 2015; De Pierri et al. 2020; Leimeister et al. 2019) are, in principle, transferable to the textual domain. This exchange broadens the analytical repertoire for text analysis (Hassani et al. 2020; Lilleberg, Zhu, and Zhang 2015; Ma and Zhang 2015), opening the way for novel approaches. Currently, Large Language Models (LLMs) such as GPT (OpenAI et al. 2023; Radford and Narasimhan 2018) offer a generalist solution to text-related problems on a scale. However, their adoption raises significant concerns (Currie 2023). The epistemic complexity of deep learning severely limits interpretability and complicates its reliability assessment (Scorzato 2024). For researchers aiming to solve specific, specialized problems modeled from scratch, these challenges are magnified by limited access to proprietary training data, opaque model training processes, and constrained computational resources for domain adaptation. Consequently, a consensus is emerging around the need for more transparent, controllable, and computationally manageable alternatives that do not sacrifice analytical power (Hutson 2024).

Driven by this need, this study presents the theoretical conceptualization of Biotext, a framework that integrates Bioinformatics and Natural Language Processing (NLP) through biological sequence-inspired text processing. The validity of this general approach has been preliminarily demonstrated in a published material that

applied it to the Yoga literature (Ieger-Raittz et al. 2025). However, the present article shifts the focus from empirical application to formalization, offering a detailed conceptual and theoretical foundation for the framework. At the core of Biotext lies SWeePtex, a formal adaptation of the biological sequence vectorization method SWeeP (De Pierri et al. 2020) for textual data. The framework theoretically articulates two distinct strategies for converting text into Biological Sequence-Like (BSL) representations: AMINOcode (an amino acid-like encoding) and DNAbits (a nucleotide-like encoding).

The primary objective of this article is to present the conceptual design of the Biotext framework and its SWeePtex method. We also provide an example that applies the framework to analyze the scientific literature on thioredoxin from PubMed. This demonstration employs unsupervised machine learning and geometric analysis of the generated text vectors to identify patterns within the corpus, serving strictly as a proof of concept for the theoretical propositions. To support the proposed approach, a review is structured as follows: first, it explores the vectorization of natural language (Section 1.1), then provides a detailed theoretical explanation of the original SWeeP method (Section 1.2), and finally positions SWeePtex within the relevant specialized literature (Section 1.3).

1.1 Natural language vectorization

Natural language modeling (NLM) explores the intersection of computers and human language. The primary objective of NLM is to equip machines with the ability to understand, interpret, and generate text at a level comparable to human cognition (Jurafsky and Martin 2025). Language modeling has advanced significantly in recent decades, from traditional word-vectorization methods to sophisticated models that leverage deep neural networks and reinforcement learning paradigms.

Inverse Document Frequency (IDF), an early and foundational approach to text vectorization, is rooted in the theory of Jones (1972). The metric assigns a lower weight to terms that appear frequently across many documents in a corpus. Combined with Term Frequency (TF), it forms the TF-IDF weighting scheme, which has been extensively applied in information retrieval and TM (Robertson 2004). Despite the advent of neural methods, the core intuition behind IDF continues to influence contemporary NLP.

Recurrent Neural Networks (RNNs) (Elman 1990) enable the representation of words and sequential data by processing inputs through time-dependent hidden states. These networks preserve temporal information in continuously updated vector representations, allowing them to capture contextual and sequential relationships between words. This architecture represented a significant advance in neural language processing by enabling models to encode both lexical semantics and sequential structure.

Random-based approaches also play an important role in text vectorization.

Random Indexing, proposed by Kanerva (1994), employs randomly generated vectors to represent concepts, drawing inspiration from cognitive models. In this framework, dense vectors encode generalizations, while sparse vectors capture specific features, enabling the formation of new representations through vector combination (Kanerva 1994; Kanerva, Kristoferson, and Hols 2000). Kanerva (1994) argues that such randomness enables flexible, non-deterministic concept formation, loosely analogous to mechanisms of human cognition.

Random Mapping (Kaski 1998) reduces vector dimensionality while approximately preserving pairwise similarities, relying on the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984). When applied to the Web Self-Organizing Map (WEBSOM) system, this method projects documents into a lower space while maintaining thematic separability, producing results comparable to those obtained with Principal Component Analysis (PCA) (Samuel Kaski et al. 1998).

The Neural Language Model, introduced by Bengio et al. (2003), employs neural networks to generate distributed word representations based on contextual information. This research establishes a foundational approach for learning continuous word vectors, thereby capturing the statistical regularities of language in a high-dimensional space. This concept directly informs the later development and formalization of word embeddings.

Word2Vec (Word to Vector) (Mikolov et al. 2013) introduced the Continuous Bag-of-Words (CBOW) and Skip-gram architectures. Both models learn distributed word representations from local context windows, but with different objectives: CBOW predicts a target word from surrounding context, while Skip-gram predicts context words from a target word. These methods significantly improved the efficiency and semantic quality of learned embeddings.

FastText (Bojanowski et al. 2016) extends Word2Vec by incorporating subword information through character n-grams. This design improves robustness when representing rare or out-of-vocabulary words, while retaining efficient gradient-based training of vector representations.

GPT (Generative Pre-trained Transformer) (Radford and Narasimhan 2018) represents text using a unidirectional self-attention mechanism within a Transformer-based architecture. During processing, the model attends only to preceding tokens in the sequence. This constraint produces contextual vector representations optimized for autoregressive language modeling tasks, such as text generation.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018) is built upon a stack of standard Transformer encoders. Its primary pretraining objective, Masked Language Modeling (MLM), enables token representations to be conditioned on their entire surrounding context, both left and right, simultaneously. This explicit bidirectionality yields contextual representations that integrate information from the full input sequence, providing a holistic view of context that is particularly advantageous for text analysis and understanding tasks in which meaning emerges from the complete sentence.

T5 (Text-to-Text Transfer Transformer) (Raffel et al. 2019) proposes a unified framework for NLP by casting all tasks into a text-to-text format. In this architecture, a Transformer encoder maps input text to contextual latent representations, while a decoder generates the output sequence. Pre-training relies on a span-corruption objective, in which contiguous token spans are masked and reconstructed from large-scale corpora. The resulting representations are adapted to diverse downstream tasks through fine-tuning, primarily by modifying the task-specific instruction text.

DeepSeek-R1 (DeepSeek Reasoner 1) (DeepSeek-AI et al. 2025) employs a representation learning through a multi-stage reinforcement learning framework. Its distilled variants produce more compact vector representations by transferring learned behaviors to smaller architectures. In this teacher-student paradigm, DeepSeek-R1 guides student models, such as Qwen and LLaMA, to approximate its internal reasoning patterns and output distributions, thereby integrating its learned representations.

Supplementary Table S1 presents a chronological sequence of the methodologies described and the progress made for each. The table highlights the evolution of key concepts and methods in linguistic modeling, ranging from early approaches. Each entry reflects progress in computational linguistics and the contributions that have shaped the field's trajectory.

1.2 Sequence vectorization with SWeeP

SWeeP (Spaced Words Projection) (De Pierri et al. 2020) is a method that efficiently represents biological sequences as compact vectors using FASTA-formatted data as input. SWeeP operates in two steps: 1) creation of a High-Dimensional Vector (HDV); and 2) pseudo-random projection of the HDV into a Lower-Dimensional Vector (LDV or SWeeP vector). The random projection is based on the concept proposed by Johnson and Lindenstrauss (Johnson and Lindenstrauss 1984), which reduces the dimensionality while preserving the comparability of information among vectors.

SWeeP performs well even for vectorizing long string patterns that are difficult to handle directly in the FASTA string format. The method is an efficient, quick, and effortless way of adapting sequences to machine learning tasks. SWeeP is efficient for sequence comparisons, as demonstrated in previous studies (Silva Filho et al. 2021; De Pierri et al. 2020; Perico et al. 2022; Raittz et al. 2021).

1.3 Theoretical context

The evolution of language modeling has progressed from classic statistical methods, such as IDF (Jones 1972), to complex systems, including GPT (Radford and Narasimhan 2018), BERT (Devlin et al. 2018), T5 (Raffel et al. 2019), and DeepSeek-

R1 (DeepSeek-AI et al. 2025) (Supplementary Table S1). This trajectory reveals a clear trend toward increasing complexity in capturing intricate semantic patterns.

Within this diverse landscape, SWeePtex emerges as a distinct approach rooted in a random-projection representation. It aligns well with techniques such as Random Indexing (Kanerva 1994) and Random Mapping (Kaski 1998; Samuel Kaski et al. 1998). These methods show that randomized projections can efficiently preserve semantic relationships through dimensionality reduction, providing a computationally lightweight alternative to more resource-intensive dense embeddings.

Early neural models, including recurrent networks (Elman 1990) and the Neural Language Model (Bengio et al. 2003), are groundbreaking in pioneering learned distributed representations, yet they still carried significant computational overhead. In contrast, SWeePtex capitalizes on the inherent simplicity of random projection, effectively bypassing the extensive training phases and parameter fine-tuning required by later models, such as Word2Vec (Mikolov et al. 2013) and FastText (Bojanowski et al. 2016). Although the Transformer architecture introduced powerful attention mechanisms for deep contextual modeling, it also exponentially increased computational demands. Even complex architectures such as DeepSeek-R1, which aim to distill knowledge into compact representations, still reflect the idea of random-based dimensionality reduction.

SWeePtex has carved out a unique niche in this spectrum. It combines the efficiency of early random projection methods with a BSL encoding layer. Although it does not employ attention mechanisms or iterative parameter optimization, SWeePtex incorporates contextual information holistically at the sentence level. In this respect, it can be theoretically related to contextual language models such as BERT, which derive meaning from the entire sentence, albeit through deep bidirectional attention.

While model distillation, as exemplified by DeepSeek-R1 (DeepSeek-AI et al. 2025), projects knowledge iteratively from a teacher to a student model, SWeePtex operates under a fundamentally different paradigm by replacing this costly training process with efficient, non-iterative random projection. Importantly, SWeePtex explicitly embraces randomness as a core design feature, aligning itself with the neural random concept championed by Kanerva (1994).

2 Method

The methodological approach of this study is structured into three phases. First, we establish the theoretical framework and design specifications for the Biotext system, defining its core components and their functional relationships. Building upon this foundation, we next implement the framework as a functional Python package. Finally, we conduct an experiment to demonstrate the system's utility through a usage example.

2.1 Theoretical design

The Biotext framework integrates a suite of techniques. It includes the two BSL encoding schemes: AMINOcode and DNAbits. The framework also uses the SWeePtex method, which combines BSL encodings with the SWeeP algorithm. Additionally, it employs a contextual embedding strategy derived from SWeePtex. Each component is described with intuitive explanations and formal mathematical definitions, and its computational complexity is analyzed theoretically.

2.2 Implementation approach

The Biotext framework is implemented as a Python 3 package (Van Rossum and Foundation 2026). The implementation includes modules for BSL encoding and SWeePtex vectorization, providing a user-friendly interface for researchers. A pipeline script written in the same language automates the usage example experiment.

In the Biotext package, the NumPy library (Harris et al. 2020) is used for vector manipulation procedures. The results are made publicly available through the PyPI repository. Biopython (Cock et al. 2009) is used to manipulate FASTA files. The Scikit-learn library (Pedregosa et al. 2011) is used to implement machine learning in the experimental script. Graphs are generated using Matplotlib (Hunter 2007) and Wordcloud (Mueller 2026), while scatterplot refinement is performed using the adjustText (Flyamer et al. 2024) package. The Natural Language Toolkit (NLTK) (Bird, Klein, and Loper 2009) stopwords list is considered when necessary. The dataset is drawn from the PubMed database (Canese and Weis 2002).

2.3 Usage example: “thioredoxin”

An experiment demonstrates SWeePtex’s capabilities for integrating machine learning and geometric methods in text analysis and visualization. It presents a complete, executable pipeline for analyzing MEDLINE abstracts to extract biomedical information. This usage example serves as a proof-of-concept demonstration; its primary purpose is to illustrate the practical application of the theoretically conceptualized Biotext framework. Therefore, methodological choices regarding encoding schemes, parameter settings, and analytical techniques are presented as illustrative examples rather than optimized solutions. The complete, modifiable source code is publicly available, allowing researchers to explore alternative configurations according to their specific requirements. The Supplementary Box S1 contains a demonstration of the usage example pipeline using fictitious data.

The topic of “thioredoxin” serves as a practical demonstration case, chosen because it reflects the research group’s genuine interest arising from a previous study (Rubel et al. 2016). This selection demonstrates the applicability of the protocol to real-world research questions.

The dataset is a MEDLINE (PubMed format) file constructed from a PubMed search (12 April 2025) for the topic "thioredoxin", comprising 14,984 entries (documents). The corpus is converted to lowercase and tokenized to ensure consistency in subsequent analyses. The search details are in the Supplementary Table S6.

Term Frequency-Inverse Document Frequency (TF-IDF) norms are computed for all words in the corpus, with a threshold of greater than or equal to 0.5 applied to select domain-relevant vocabulary.

Semantic representations of words and documents are generated using Biotext embeddings with AMINOcode to convert text into BSL format and SWeePTex to vectorize. Principal Component Analysis (PCA) is applied to reduce the dimensionality of both word and document embeddings, retaining the first 50 principal components.

For unsupervised word-level analysis, the optimal number of clusters is determined using the elbow method, and clustering is performed using k-means.

At the document level, the most salient terms are extracted using a TF-IDF cutoff and stop-word filtering. The elbow criterion determines the optimal number of clusters, and k-means is applied to group documents thematically. The document clusters are represented by items selected using two approaches: (1) centroid-proximate, where documents closest to the geometric center of the cluster (centroid) are chosen for their typicality and representativeness of the core themes, and (2) randomly sampled, where documents are selected arbitrarily to capture a greater variability within the group.

The pipeline illustrated most of the resources available through the Biotext package, as well as other potential analyses that can be performed on the generated data. The results are presented in multiple formats, including dimensional-reduction visualizations, cluster-annotation plots, frequency-based word clouds, and structured tabular data summarizing key findings at both the word and document levels.

Finally, a manual analysis of document clusters is performed using the generated material to identify the pattern recognized by unsupervised machine learning.

3 Result

The results are presented in three parts. First, we detail the theoretical approaches and design of the Biotext framework, which translates natural language into Biological Sequence-Like (BSL) formats and projects these sequences into a vector space. Second, we describe its practical implementation in an open-source Python package. Finally, we demonstrate its application in a usage example from a biomedical corpus.

3.1 Theoretical approaches and design

The Biotext framework is conceptually built upon two core BSL encoding methods, AMINOcode and DNAbits, which convert natural language into Biological Sequence-Like (BSL) formats. These encoded sequences are designed to be compatible with SWeeP vectorization through the SWeePtex method. The resulting vectors enable downstream tasks, including semantic embedding and machine learning applications.

3.1.1 AMINOcode

AMINOcode is a method for encoding natural language text into a format based on amino acid sequence representation in FASTA format. It provides two encoding variants: reduced and detailed. In the reduced variant, the resulting sequence is compact, but it loses information that cannot be recovered during decoding. The detailed variant increases the sequence length while retaining information for additional characters, such as numbers and significant punctuation marks. Supplementary Table S2 specifies the character-substitution rules, and Supplementary Table S3 illustrates example encodings with both variants, showing the loss of information upon decoding. Notably, AMINOcode does not preserve letter case or characters that are not in its substitution dictionary.

Let a text T be a sequence of characters $T = (c_1, c_2, \ldots, c_n)$, where each c_i belongs to an admissible symbol set $\mathcal{C}$. AMINOcode is defined by an encoding function $\psi : \mathcal{C} \to \Sigma_{\text{aa}}^{+}$ that maps each character to a non-empty string over the 20-symbol amino acid alphabet Σ_{aa}. The function operates in two distinct modes. In reduced mode, the domain is restricted to a subset $\mathcal{C}' \subset \mathcal{C}$, and multiple characters are mapped to shared amino-acid codewords of fixed length, resulting in a space-efficient but non-injective encoding in which distinctions such as digits, punctuation, letter case, and unsupported symbols are irreversibly lost. In detailed mode, the mapping is expanded to cover a larger subset of $\mathcal{C}$, including digits and common punctuation, by assigning distinct amino-acid codewords of variable length to each supported character, thereby preserving symbol-level distinctions at the cost of increased sequence length. The encoded biological sequence is obtained by concatenation:

$$S_{\text{aa}}(T) = \psi(c_1) \| \psi(c_2) \| \ldots \| \psi(c_n), \quad \psi(c_i) \in \Sigma_{\text{aa}}^{+}.$$

The decoding employs a partial inverse $\psi^{-1} : \Sigma_{\text{aa}}^{L} \rightharpoonup \mathcal{C}$ defined only for codewords present in the encoding dictionary. Consequently, decoding cannot recover the collapsed distinctions during encoding, reflecting the inherent trade-off between sequence compactness and information retention.

The AMINOcode exhibits linear time complexity $O(n)$, where n is the number of characters in the input text, since each character is processed via a constant-time dictionary lookup. Memory usage is determined by the encoding expansion factor,

which varies according to the selected mode. In reduced mode, most alphabetic characters are encoded as a single amino acid, while digits and punctuation require two amino acids, yielding a variable expansion per character of either 1 or 2 symbols. In expanded mode, characters are encoded using between 1 and 3 symbols, producing a more expressive but less compact representation. In both cases, the expansion is linear in the input size, and the overall space complexity remains $O(n)$, with the dictionary storage overhead being negligible.

3.1.2 DNAbits

DNAbits encodes natural-language text into a DNA-sequence representation in FASTA format. The output sequence is longer than that produced by AMINOcode; however, it preserves all information from the original American Standard Code for Information Interchange (ASCII), as detailed in Supplementary Table S4. The method operates by converting each character into its 8-bit ASCII binary representation in least-significant-bit-first order, splitting each byte into four consecutive 2-bit pairs, and mapping each pair to a nucleotide: $00 \to A$, $10 \to C$, $01 \to G$, and $11 \to T$. For example, the character "a" (ASCII value 97) has the binary representation `10000110` (in least-significant-bit–first order), which is split as `10-00-01-10`, yielding the nucleotide sequence "CAGC". Optionally, the resulting DNA sequence can be translated into amino acids using the standard genetic code across the three forward reading frames, and the resulting amino acid sequences are concatenated to yield an alternative protein-like representation.

Given a text $T = (c_1, c_2, \ldots, c_n)$ with each character $c_i \in \mathcal{C}$ (ASCII set). Denote by $b(c_i) \in \{0,1\}^8$ the 8-bit binary expansion of the ASCII code of c_i. This byte is divided into four consecutive 2-bit pairs $p_1p_2p_3p_4$. A fixed mapping $\eta : \{0,1\}^2 \to \Sigma_{\text{dna}}$ is applied, where $\Sigma_{\text{dna}} = \{A, C, G, T\}$ and

$$\eta(00) = A, \quad \eta(10) = C, \quad \eta(01) = G, \quad \eta(11) = T.$$

The DNAbits-encoded sequence of c_i is

$$\phi_{\text{dna}}(c_i) = \eta(p_1)\,\eta(p_2)\,\eta(p_3)\,\eta(p_4) \in \Sigma_{\text{dna}}^4,$$

and the full encoded text becomes

$$S_{\text{dna}}(T) = \phi_{\text{dna}}(c_1) \| \phi_{\text{dna}}(c_2) \| \ldots \| \phi_{\text{dna}}(c_n) \in \Sigma_{\text{dna}}^{4n}.$$

The optional amino-acid translation uses the genetic code $\tau : \Sigma_{\text{dna}}^3 \to \Sigma_{\text{aa}}$ (where Σ_{aa} is the amino-acid alphabet of 20 symbols, possibly including a stop symbol). Translating $S_{\text{dna}}(T)$ into the three forward reading frames and concatenating yields

$$S_{\text{aa-alt}}(T) = \tau\big(S_{\text{dna}}(T)[1:]\big) \| \tau\big(S_{\text{dna}}(T)[2:]\big) \| \tau\big(S_{\text{dna}}(T)[3:]\big),$$

where $S_{\text{dna}}(T)[k:]$ denotes the subsequence starting at position k and is taken in non-overlapping triplets. This two-step encoding ensures complete reversibility to the original ASCII text from $S_{\text{dna}}(T)$. At the same time, amino-acid-like translation from DNAbits sequences can also be applied to compactification.

The DNAbits operates with linear time complexity $O(n)$, where n is the length of the input text. The space complexity of the output sequence is $O(n)$, with an exact expansion to $4n$ characters. This deterministic length arises from mapping each 8-bit ASCII character to exactly four nucleotide symbols. The optional subsequent step of three-frame amino acid translation yields a total length of approximately $4n$ amino acids across all reading frames when the translated sequences are concatenated, as each reading frame produces on the order of $4n/3$ amino acids and the concatenation of the three frames results in $(4n/3) \times 3$, with minor deviations depending on how residual nucleotides at sequence ends are handled.

3.1.3 SWeePtex

SWeePtex integrates BSL encoding (AMINOcode or DNAbits) with the Spaced Words Projection (SWeeP) method to transform texts into numerical vectors, independent of their original length. After a text is encoded into a biological sequence using either AMINOcode or DNAbits, the SWeeP algorithm is applied directly to the resulting sequence without further preprocessing.

SWeeP operates in two steps for each spaced-word mask: first, it constructs a high-dimensional vector (HDV) by counting spaced-word patterns specific to that mask; second, it projects each HDV into a lower-dimensional space using a random projection matrix, following the Johnson-Lindenstrauss lemma. For a set of masks, the resulting lower-dimensional vectors are concatenated to form the final representation. The combined pipeline, BSL encoding followed by SWeeP projection and concatenation, is termed SWeePtex. The resulting vectors have a uniform dimensionality, enabling their immediate use in machine-learning workflows, PCA, graph construction, matrix operations, text-similarity assessment, semantic analysis, and other applications that benefit from vector-space representations of textual data.

Consider a text T encoded as BSL $E(T) \in \Sigma^*$ by AMINOcode or DNAbits, where Σ is the corresponding alphabet (Σ_{aa} or Σ_{dna}). Using a set of spaced-word masks $M \subset \{0,1\}^k$ (length k, weight H), a hash function $h : \Sigma^H \to \{1,\ldots,d\}$ maps each pattern to an index in an implicit d-dimensional space, with $d = |\Sigma|^H$. For a given mask $m \in M$, an HDV $u^{(m)}(T) \in \mathbb{R}^d$ records pattern frequencies specific to that mask:

$$u_j^{(m)}(T) = \sum_{p=1}^{|E(T)|-k+1} \mathbb{I}\big[\, h(\text{extract}_m(E(T),p)) = j \,\big], \quad j = 1,\ldots,d,$$

where j indexes the coordinates of the high-dimensional vector and corresponds to a specific hashed spaced-word pattern. Here, $S = E(T) \in \Sigma^*$ denotes the BSL-

encoded sequence of the text, and $\text{extract}_m(S, p)$ returns the subsequence of length H obtained from S at positions where the mask m has value 1, starting at position p.

For each mask m, a random projection matrix $A^{(m)} \in \mathbb{R}^{m' \times d}$ projects the mask-specific HDV into a lower-dimensional vector:

$$v^{(m)}(T) = A^{(m)}\, u^{(m)}(T) \in \mathbb{R}^{m'}.$$

The final SWeePtex vector is obtained by concatenating the projected vectors from all masks:

$$v_{\text{SWeePtex}}(T) = \Big[v^{(1)}(T) \| v^{(2)}(T) \| \ldots \| v^{(|M|)}(T) \Big] \in \mathbb{R}^{m' \cdot |M|}.$$

By the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984), a set of N vectors can be embedded into

$$m' = O(\varepsilon^{-2} \log N)$$

dimensions while preserving all pairwise Euclidean distances within a factor of $1 \pm \varepsilon$ with high probability. For each spaced-word mask m, the corresponding random projection approximately preserves the Euclidean distances between the high-dimensional vectors (HDVs) derived under that mask, maintaining their local similarity structure in the projected space. The concatenated representation, formed by aggregating multiple mask-specific projections, increases robustness and expressiveness. Consequently, SWeePtex yields a computationally efficient, dimension-uniform representation that bridges symbolic text patterns with geometric vector-space methods, enabling scalable similarity computations.

Two main stages determine the computational complexity and memory requirements of SWeePtex: HDV construction and random projection. For a sequence of length L_s, HDV construction requires $O\big(|M| \cdot (L_s - k + 1) \cdot H\big)$ operations to extract and hash spaced patterns for each mask $m \in M$, where k is the mask length and H is the mask weight. Each mask produces an independent HDV, which is subsequently multiplied by a random projection matrix to obtain an LDV. While the nominal cost of this projection is $O(m' \cdot d)$, with $d = |\Sigma|^H$, the HDVs are sparse in practice, containing only $\text{nnz} \ll d$ non-zero entries. Exploiting this sparsity reduces the effective projection cost to $O(m' \cdot \text{nnz})$ per mask. The total computational complexity per sequence is therefore

$$O\Big(|M| \cdot \big(L_s \cdot H + m' \cdot \text{nnz}\big)\Big),$$

which scales linearly with the sequence length and the number of masks. Scalability is ensured when the projection dimension m' is chosen logarithmically with respect to the number of sequences, as motivated by the Johnson-Lindenstrauss lemma.

Memory usage involves three primary components: the encoded sequence ($O(L_s)$), the sparse HDVs during processing ($O(|M| \cdot \text{nnz})$), and the final fixed-dimension output vectors ($O(m' \cdot |M|)$ per document). For a corpus of N documents, the total storage of all vectors is $O(N \cdot m' \cdot |M|)$, which remains manageable due to the logarithmic scaling $m' = O(\log N)$. Peak memory occurs transiently during HDV construction, but is optimized through streaming processing and sparse data structures.

3.1.4 SWeePtex embedding

The SWeePtex embedding enables the creation of context-aware word and document embeddings by leveraging bidirectional relationships within a corpus. The process operates on a corpus $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$ of documents. Embeddings are constructed through a two-phase, iterative procedure that first computes preliminary document vectors and then refines them via word-document co-occurrence statistics. Figure 1 illustrates the complete workflow. The Supplementary Box S2 shows a demonstration of the process.

Formally, let $\mathcal{W} = \{w_1, w_2, \ldots, w_V\}$ be the vocabulary extracted from $\mathcal{D}$. For each document D_i, a preliminary SWeePtex vector $d_i^{(0)} \in \mathbb{R}^m$ is obtained directly from the raw text using the standard SWeePtex pipeline (BSL encoding followed by SWeeP). These preliminary vectors correspond to the $v_{\text{SWeePtex}}(T)$ representations described in the previous section, with dimensionality $m = m' \cdot |M|$, and serve as initial context representations.

Word embeddings are then derived as the average of the preliminary vectors of all documents in which the word appears. Let $\mathcal{D}(w)$ denote the set of documents that contain the word w. The word embedding $e_w \in \mathbb{R}^m$ is computed as:

$$e_w = \frac{1}{|\mathcal{D}(w)|} \sum_{D_i \in \mathcal{D}(w)} d_i^{(0)}.$$

Subsequently, the final document embeddings are obtained by averaging the embeddings of all words contained in the document. For document D_i with word sequence $(w_{i1}, w_{i2}, \ldots, w_{iL_i})$, the refined document embedding $d_i \in \mathbb{R}^m$ is:

$$d_i = \frac{1}{L_i} \sum_{j=1}^{L_i} e_{w_{ij}}.$$

This bidirectional averaging procedure effectively propagates contextual information across the corpus: document vectors inform word representations, and word vectors, in turn, refine document representations. The resulting embeddings capture both local lexical statistics and global corpus-level co-occurrence patterns, making them suitable for downstream tasks such as semantic similarity measurement, document classification, and information retrieval.

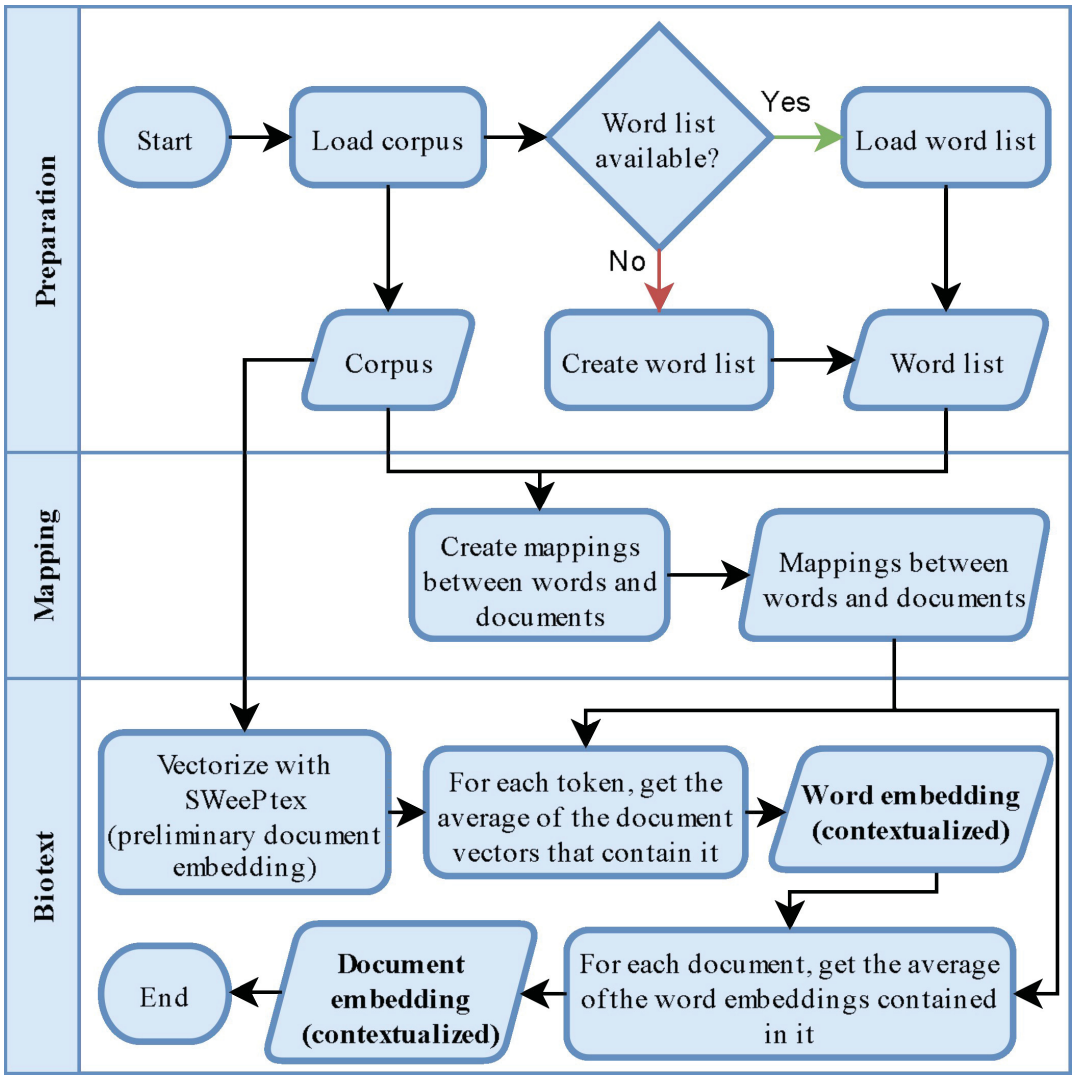


Figure 1: SWeePtex embedding flowchart. This figure illustrates the SWeePtex embedding technique. It involves three key steps: (1) loading the textual corpus, (2) creating a bidirectional mapping between words and documents, and (3) computing embeddings for both words and documents through vector averaging operations.

The actual embedding program checks whether a preexisting word list is available. If such a list exists, it is loaded directly for use. If not, the program creates a new word list by extracting all unique terms from the corpus. Next, the program begins mapping, establishing bidirectional relationships between words and documents.

The SWeePtex embedding algorithm proceeds through three main computational stages. The overall time complexity scales as $O(N \cdot (C_{\text{SWeePtex}} + \bar{L}_d \cdot m))$, where N is the number of documents, C_{SWeePtex} the cost per-document of the SWeePtex base vectorization, $\bar{L}_d$ the average document length in words, and m the embedding dimension. First, preliminary SWeePtex vectors are generated for all documents,

which require $O(N \cdot C_{\text{SWeePtex}})$ operations. The word embeddings are constructed by aggregating, through an inverted index, the document vectors associated with each term in the vocabulary $\mathcal{W}$ (size V), which takes $O(N \cdot \bar{L}_d \cdot m)$ in the worst case. Finally, refined document embeddings are obtained by averaging the word embeddings contained in each document, which also costs $O(N \cdot \bar{L}_d \cdot m)$.

The space complexity is determined by the storage of three vector sets and the inverted index. Memory must hold preliminary document vectors ($O(N \cdot m)$), word embeddings ($O(V \cdot m)$), and final document embeddings ($O(N \cdot m)$). Additionally, inverted index mapping words to documents requires $O(N \cdot \bar{L}_d)$ space. Consequently, the total memory complexity is $O((N + V) \cdot m + N \cdot \bar{L}_d)$.

3.2 Python package implementation

The Biotext Python package is freely available for installation via PyPI[1]. This implementation provides researchers with full access to the SWeePtex methodology in a user-friendly format. The package handles all stages of the analysis, from raw text input to final output visualization.

Table 1 summarizes the core modules of the Biotext package, accompanied by working examples. These demonstrate BSL encoding (with both the AMINOcode and DNAbits options) and the construction of a vector space.

The package includes the exact executable pipeline used for the thioredoxin experiment. The documentation provides detailed guidance for running the complete experiment. Beyond reproduction, the architecture enables flexible modifications for new applications. Researchers can adjust encoding schemes, modify projection dimensions, or implement custom similarity metrics to optimize their analysis. This modular design supports both immediate out-of-the-box use and extensive customization for specialized needs.

3.3 Thioredoxin usage example result

The implemented TM pipeline generated multiple insightful visualizations and quantitative outputs, revealing patterns across the 14,984 MEDLINE abstracts analyzed. The complete execution takes approximately 1 hour and 22 minutes on a personal computer (hardware details in the Supplementary Table S5), with approximately 16 minutes dedicated to Biotext embedding (words and documents) and the remaining time allocated to pre- and post-processing tasks.

All experimental material, including the complete pipeline implementation, is publicly available through the Zenodo repository[2], comprising the main execution script, supporting utility modules, configuration files, and all output results utilized

[1] `https://pypi.org/p/biotext`

[2] `https://doi.org/10.5281/zenodo.18370817`

Table 1: Biotext Python package summary. This table summarizes the Biotext modules designed for generating Bioinformatics-inspired text encodings and embeddings at both the word and document levels.

Module	Description	Example
aminocode	Encodes text using amino acid-like representation.	`encoded = aminocode.encode_string( "Hello world!" ) # Output: # 'HYELLYQYSYWYQRLDYPW'`
dnabits	Encodes text using DNA-like representation.	`encoded = dnabits.encode_string( "Hello world!" ) # Output: # 'AGACCCGCATGCATGCTTGCAAGAT #  CTCTTGCGATCATGCACGCCAGA'`
sweeptex	Generates document vectors using the SWeePtex algorithm.	`embeddings = sweeptex( ["Text 1", "Text 2"], emb_size=1200 ) # Output array shape: # (2, 1200)`
sweeptex_emb	Generates word and document embeddings via a processing pipeline.	`results = biotext_emb( ["First doc", "Second doc"], return_doc_emb=True, return_word_emb=True ) # Output keys: # - Document embeddings #   results['doc_emb'] # - Word embeddings #   results['word_emb']`

in this study, thereby enabling full reproducibility of our findings while also providing researchers with adaptable components for future applications.

An elbow method analysis determines the optimal number of clusters for words and documents (Supplementary Figure S1). For words, the point of diminishing returns indicates 7 as the optimal number. For documents, the corresponding optimal number is 3.

The PCA of the word and document embeddings reveals their organization into distinct clusters (Figure 2). For word embeddings (Panel A), clusters are identified by their most representative terms, which are determined by the proximity of their centroids. Similarly, document embeddings (Panel B) are grouped, with each cluster labeled with the most representative TF-IDF word from its documents. Furthermore, the semantic neighborhoods of "leukemia" (Panel C) and "pregnancy" (Panel D) are visualized, illustrating the conceptual proximity of related terms within the embedding space.

For each target word ("leukemia" and "pregnancy"), the titles and PubMed IDs (PMID) of their most closely associated documents are detailed in the Supplemen-

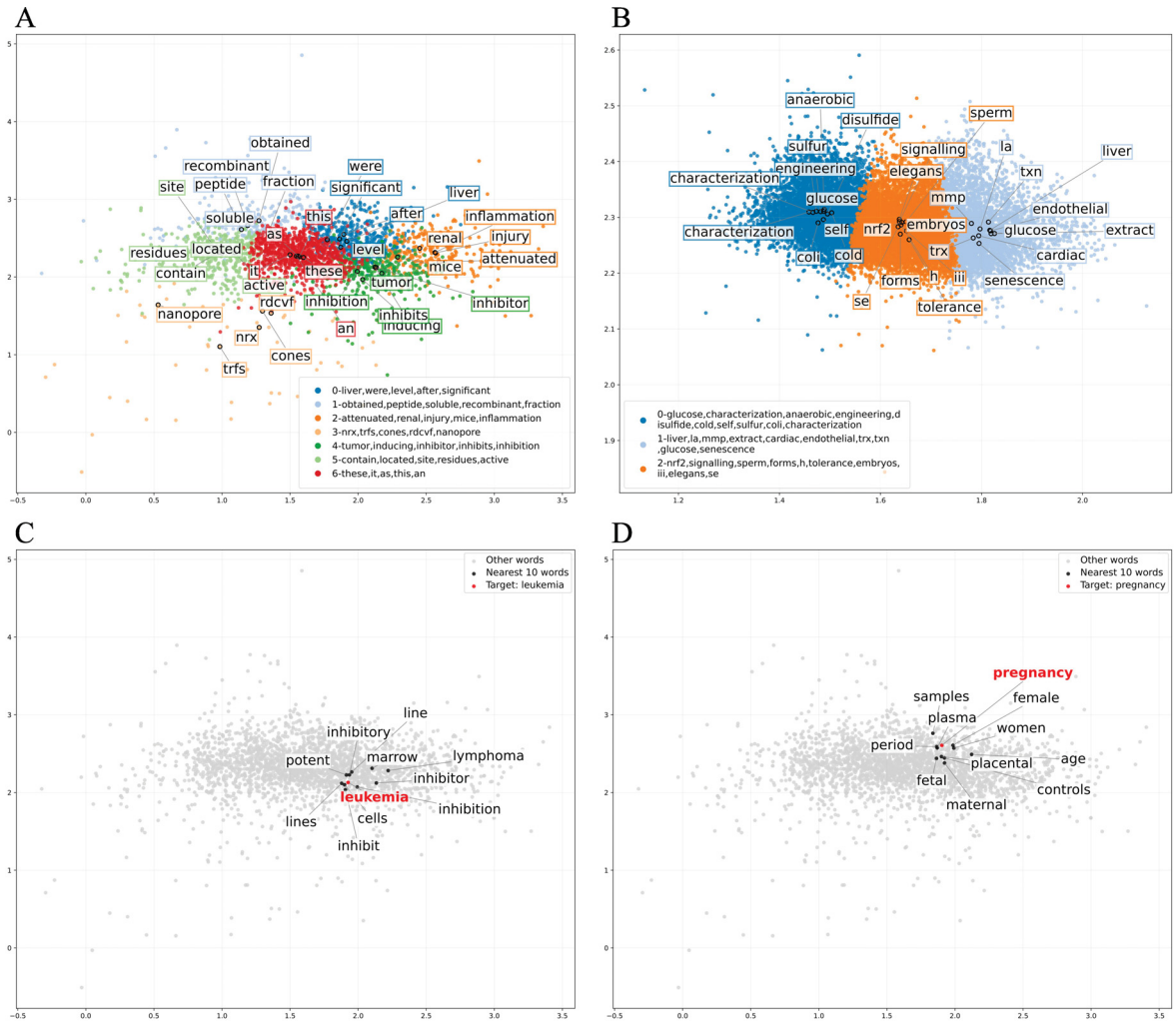


Figure 2: Thioredoxin usage example: scatter plot visualizations. This figure illustrates the results of the usage example through two-dimensional scatter plots of PCA-reduced embedding data: (A) Word embeddings are clustered by k-means and annotated with five representative terms per cluster, defined by their proximity to the centroid; (B) Document embeddings are displayed, colored by cluster, and labeled with TF-IDF-derived words representing the documents closest to each centroid; (C) The neighborhood of "leukemia" shows its ten closest words plus the target word; and (D) The analogous neighborhood for "pregnancy" displays its ten closest words.

tary Table S7. Word clouds also provide cluster characterization for each document group (Supplementary Figure S2). The 30 most frequent representative terms for each document group are summarized in Supplementary Table S8. Supplementary Table S9 provides an analysis of the groupings of documents based on the cited material. At the same time, the Supplementary Table S10 presents examples of documents for each cluster, selected using proximate centroid and random sampling, along with interpretations of their cluster membership.

4 Discussion

This discussion interprets the results by examining the underlying conceptualization, computational implications, and specific insights from the thioredoxin usage example. We explore how the Biotext framework bridges Bioinformatics methods with natural language processing, analyze its efficiency profile compared to contemporary models, and evaluate its practical utility for structuring and exploring biomedical literature. Together, these perspectives contextualize SWeePtex as a complementary approach that prioritizes transparency, scalability, and immediate applicability in text representation.

4.1 Conceptualization

Biotext with SWeePtex demonstrates how Bioinformatics methods can effectively address key challenges in NLP by bridging symbolic sequence analysis with text representation. Although the BSL format has been used in third-party implementations (Araujo et al. 2022), SWeePtex, to our knowledge, represents the first implementation specifically designed for text vectorization within this paradigm. Provides interpretable and content-addressable representations that enable direct vector manipulation for literature exploration, thereby establishing continuity with randomness-based approaches to linguistic representation.

The methodology is aligned with established strategies in cognitive and computational linguistics that employ random projections for representation. Kanerva (1994) demonstrated how sparse distributed memory and random indexing can reflect human conceptual cognition, providing a theoretical foundation for high-dimensional geometric approaches to meaning. SWeePtex operationalizes this principle within a Bioinformatics framework, applying spaced-word hashing and random projection to create text embeddings.

The semantic construction in SWeePtex embedding follows a theoretically grounded approach based on vector averaging. Primary SWeeP vectors initially capture basic lexical information from individual texts, while semantic representation emerges through a contextualization process mediated by vector averages. The contextual vector for each word is calculated as the average of the vectors from all the documents in which it appears. This allows words to transcend isolated representation and incorporate shared semantic context across documents, following the distributional principle that meaning arises from co-occurrence patterns (Firth 1957). Subsequently, each document's vector is obtained by averaging the contextual vectors of all its words, thereby semantically synthesizing the document from its already contextualized terms. This is a consolidated technique for composing document representations (Mitchell and Lapata 2010).

This design favors document-level similarity tasks, where computational efficiency and straightforward implementation are valued over explicit modeling of word

order or complex syntactic structures. This aspect could be addressed in future work through strategies such as positional encoding or syntax-aware pattern extraction.

Although transformer-based models have dominated recent NLP research, they represent a fundamentally different paradigm based on learned, contextually modulated representations. The approach implemented in SWeePtex offers a complementary perspective: instead of teaching representations via gradient-based optimization over massive datasets, it constructs them via deterministic hashing, random projection, and theoretically grounded averaging operations. This distinction highlights an alternative pathway in representation learning that prioritizes transparency, reproducibility, and immediate applicability.

Randomized approaches and neural network methods are not mutually exclusive but rather complementary research directions. Patterns identified through random exploration can provide initialization points or constraints for neural architectures, while learned representations can inform the design of more effective hashing and projection schemes. This synergistic integration represents a promising avenue for future development, potentially yielding hybrid systems that combine the efficiency and interpretability of randomized methods with the representational power of learned models.

4.2 Computational Complexity

The computational demands of text processing methods reveal a fundamental trade-off between representational power and operational efficiency. SWeePtex operates in a deterministic, near-linear complexity regime, with a per-document processing cost that does not scale with corpus size.

In contrast, transformer-based language models exhibit quadratic time and memory complexity in the self-attention mechanism, as characterized in surveys of efficient Transformers (Tay et al. 2022). Introduced in the original Transformer architecture (Vaswani et al. 2017), self-attention reflects a profound architectural commitment to learning dense, contextually modulated representations through gradient-based optimization over massive datasets, which are powerful but computationally intensive to produce and opaque to direct interpretation.

SWeePtex circumvents this trade-off by adopting a randomization-based, geometry-first approach inspired by Bioinformatics. Its efficiency comes from avoiding the learning process altogether. Instead, it employs hashing and random projection to map textual patterns into a fixed-dimensional space, preserving similarity in accordance with the Johnson-Lindenstrauss lemma. Consequently, SWeePtex excels in scenarios where the cost, delay, or opacity of the deep learning model deployment is prohibitive. These applications include rapid exploratory analysis, dynamic or domain-specific corpora lacking pre-training data, and tasks requiring interpretable feature attributions.

Recent efforts to mitigate the computational burden of transformers acknowledge

this fundamental challenge. Strategies such as model distillation (DeepSeek-AI et al. 2025), sparse attention mechanisms (Child et al. 2019), and efficient architectural variants such as Linformer (Wang et al. 2020) and Longformer (Beltagy, Peters, and Cohan 2020) all seek to address the quadratic-complexity constraint. SWeePtex represents a more radical departure from this paradigm. Rather than approximating or optimizing the transformer architecture, it replaces the learning-centric approach entirely with a lightweight, similarity-preserving embedding that is intrinsically efficient and explainable. This positions SWeePtex not as a competitor to artificial neural methods but as a complementary strategy that expands the performance-efficiency frontier for tasks where semantic nuance can be partly traded for speed, transparency, and minimal infrastructure.

The methodological choice between these approaches ultimately depends on aligning technical assumptions with application constraints. SWeePtex demonstrates that Bioinformatics-inspired randomized methods can deliver practical, scalable text vectorization for a meaningful class of problems. By offering an interpretable alternative, it enriches the NLP toolkit and provides researchers with greater flexibility when designing text-processing pipelines. However, although theoretical complexity analysis positions SWeePtex favorably against quadratic-scaling transformer models, an empirical asymptotic characterization remains a necessary step for future projects.

4.3 Thioredoxin insights

The thioredoxin usage example demonstrates SWeePtex's ability to handle large data volumes and extract relevant contextual information. The analysis is performed on a substantial collection of documents – a volume impractical for manual evaluation –, and the results are human-interpretable.

The pipeline generates word clusters that reveal an intuitive organization of biomedical terminology, as shown in Figure 2A with a specificity gradient. Generic functional words (e.g. "these", "it", and "as" in Cluster 6) are centrally located in the vectorial space, while more specialized terms radiate outwards. Clusters 0 through 5 each represent distinct biomedical subdomains, showing how the embedding naturally groups related concepts without explicit supervision. For example, Cluster 4 ("tumor", "inhibitor") includes terminology related to cancer drug discovery, while Cluster 2 ("renal", "injury", "inflammation") reflects preclinical research on kidney disease. Highly specialized terms, such as those in Cluster 3 ("nrx", "nanopore"), appear farthest from the center, occupying more specific semantic spaces.

The SWeePtex-generated document clusters (Supplementary Table S9) provide an analytical perspective on the composition of the literature related to thioredoxin. Computational pattern recognition identified three primary research domains: (1) bacterial redox systems and biotechnological applications, (2) mechanisms of oxidative stress in metabolic and degenerative pathologies, and (3) eukaryotic re-

dox biology and developmental systems. The effectiveness of the framework in structuring this literature is evident from the combination of the examination of centroid-associated terms (Figure 2B) and frequency-based word clouds (Supplementary Figure S2). This SWeePtex-based analysis offers a method for mapping current thioredoxin research, while acknowledging that other clustering approaches may reveal alternative dimensions.

The vectorial neighborhoods of "leukemia" (Figure 2C) and "pregnancy" (Figure 2D) illustrate the approach's capability to capture related concepts. For "leukemia", terms like "inhibitory", "lymphoma", and "marrow" are identified, while "pregnancy" is associated with terms such as "placental", "maternal", and "fetal". This semantic understanding extends to recognizing synonyms and terminological variations, even without exact character matching.

Supplementary Table S7 lists texts closely related to "leukemia" and "pregnancy". The system associates documents with both "leukemia" and "leukaemia" spellings, confirming that it interprets semantic relationships rather than relying on exact term matches. For "leukemia", the system retrieves studies on thioredoxin reductase inhibitors and drug resistance, whereas "pregnancy" retrieves studies on preeclampsia and gestational diabetes. These outcomes show the system's utility in biomedical text analysis, as it extracts contextually relevant literature without relying on exact terms.

Finally, the thioredoxin usage example demonstrates an adaptable pipeline that can be extended to other research topics. The experimental material is publicly available, including documented scripts, enabling third-party users to modify and reimplement them with confidence.

5 Conclusion

This study demonstrates that Bioinformatics-inspired methods offer a viable, transparent alternative to conventional text vectorization via the SWeePtex framework. By adapting spaced-word hashing, random projection, and contextual averaging from sequence analysis to text, SWeePtex constructs interpretable, content-addressable representations without the resource-intensive training required by traditional methods. This approach provides a computationally efficient path to domain-specific embeddings while fostering an epistemological understanding of how linguistic meaning is geometrically structured. The successful application to the thioredoxin literature, alongside the already published exploration of Yoga literature (Ieger-Raittz et al. 2025), confirms its utility for exploratory analysis in specialized domains. Significantly, this research expands the SWeeP method beyond its original Bioinformatics domain, establishing a novel, randomness-based paradigm for building linguistic representations from scratch. Ultimately, SWeePtex represents a complementary pathway in representation learning that prioritizes transparency, scalability, and immediate applicability in resource-aware

environments.

References

Araujo, José Deney et al. (July 2022). "Tucuxi-BLAST: Enabling fast and accurate record linkage of large-scale health-related administrative databases through a DNA-encoded approach". In: *PeerJ* 10, e13507. ISSN: 2167-8359. DOI: 10.7717/peerj.13507. URL: https://doi.org/10.7717/peerj.13507 (visited on 02/01/2026).

Asgari, Ehsaneddin and Mohammad R.K. Mofrad (Nov. 2015). "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics". In: *PLOS ONE* 10.11, e0141287. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0141287. URL: https://doi.org/10.1371/JOURNAL.PONE.0141287 (visited on 02/01/2026).

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer". In: DOI: 10.48550/arXiv.2004.05150. URL: http://doi.org/10.48550/arXiv.2004.05150 (visited on 02/01/2026).

Bengio, Yoshua et al. (2003). "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3, pp. 1137–1155. DOI: 10.5555/944919.944966. URL: https://dl.acm.org/doi/abs/10.5555/944919.944966 (visited on 02/01/2026).

Bird, Steven, Ewan Klein, and Edward Loper (July 2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.

Bojanowski, Piotr et al. (2016). *Enriching Word Vectors with Subword Information*. DOI: 10.48550/arXiv.1607.04606. URL: https://doi.org/10.48550/arXiv.1607.04606 (visited on 02/01/2026).

Canese, Kathi and Sarah Weis (2002). *PubMed: The Bibliographic Database*. URL: https://www.ncbi.nlm.nih.gov/books/NBK153385/pdf/Bookshelf_NBK153385.pdf (visited on 02/01/2026).

Child, Rewon et al. (2019). "Generating Long Sequences with Sparse Transformers". In: DOI: 10.48550/arXiv.1904.10509. URL: http://doi.org/10.48550/arXiv.1904.10509 (visited on 02/01/2026).

Cock, Peter J A et al. (2009). *Biopython: freely available Python tools for computational molecular biology and bioinformatics*. DOI: 10.1093/bioinformatics/btp163. URL: https://doi.org/10.1093/bioinformatics/btp163 (visited on 02/01/2026).

Currie, Geoffrey M (2023). "Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?" In: *Seminars in Nuclear Medicine* 53.5, pp. 719–730. DOI: 10.1053/j.semnuclmed.2023.04.008. URL: https://doi.org/10.1053/j.semnuclmed.2023.04.008 (visited on 02/01/2026).

De Pierri, Camilla Reginatto et al. (Jan. 2020). "SWeeP: representing large biological sequences datasets in compact vectors". In: *Scientific Reports* 10.1, p. 91. ISSN: 2045-2322. DOI: 10.1038/s41598-019-55627-4. URL: https://doi.org/10.1038/s41598-019-55627-4 (visited on 02/01/2026).

DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. DOI: 10.48550/arXiv.2501.12948. URL: https://doi.org/10.48550/arXiv.2501.12948 (visited on 02/01/2026).

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv*. DOI: 10.48550/arXiv.1810.04805. URL: https://doi.org/10.48550/arXiv.1810.04805 (visited on 02/01/2026).

Elman, Jeffrey L (1990). "Finding Structure in Time". In: *Cognitive Science* 14.2, pp. 179–211. DOI: 10.1207/s15516709cog1402_1. URL: https://doi.org/10.1207/s15516709cog1402_1 (visited on 02/01/2026).

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. London. URL: https://archive.org/details/papersinlinguist0000firt (visited on 02/01/2026).

Flyamer, Ilya et al. (Oct. 2024). "Phlya/adjustText: 1.3.0". In: *Zenodo*. DOI: 10.5281/zenodo.14019059. URL: https://doi.org/10.5281/zenodo.14019059 (visited on 02/01/2026).

Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature 2020 585:7825* 585.7825, pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (visited on 02/01/2026).

Hassani, Hossein et al. (Jan. 2020). "Text Mining in Big Data Analytics". In: *Big Data and Cognitive Computing* 4.1, p. 1. ISSN: 2504-2289. DOI: 10.3390/bdcc4010001. URL: https://doi.org/10.3390/bdcc4010001 (visited on 02/01/2026).

Hunter, John D. (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55. URL: https://doi.org/10.1109/MCSE.2007.55 (visited on 02/01/2026).

Hutson, Matthew (Sept. 2024). "Forget ChatGPT: why researchers now run small AIs on their laptops". In: *Nature* 633.8030, pp. 728–729. ISSN: 0028-0836. DOI: 10.1038/D41586-024-02998-Y. URL: https://doi.org/10.1038/D41586-024-02998-Y (visited on 02/01/2026).

Ieger-Raittz, Rosangela et al. (May 2025). "What are we learning with Yoga? Mapping the scientific literature on Yoga using a vector-text-mining approach". In: *PLOS ONE* 20.5, e0322791. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0322791. URL: https://doi.org/10.1371/JOURNAL.PONE.0322791 (visited on 02/01/2026).

Johnson, William B. and Joram Lindenstrauss (1984). "Extensions of Lipschitz mappings into a Hilbert space". In: *Contemporary Mathematics*. American Mathematical Society, pp. 189–206. DOI: 10.1090/conm/026/737400. URL: https://doi.org/10.1090/conm/026/737400 (visited on 02/01/2026).

Jones, Sparck K (1972). *A Statistical Interpretation of Term Specificity and its Application in Retrieval*. URL: https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf (visited on 02/01/2026).

Jurafsky, Daniel and James H Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. URL: https://web.stanford.edu/~jurafsky/slp3 (visited on 02/01/2026).

Kanerva, Pentti (1994). "The Spatter Code for Encoding Concepts at Many Levels". In: *ICANN '94*. London: Springer London, pp. 226–229. DOI: 10.1007/978-1-4471-2097-1_52. URL: https://doi.org/10.1007/978-1-4471-2097-1_52 (visited on 02/01/2026).

Kanerva, Pentti, Jan Kristoferson, and Anders Hols (2000). *Random Indexing of Text Samples for Latent Semantic Analysis*. URL: https://escholarship.org/uc/item/5644k0w6 (visited on 02/01/2026).

Kaski, S (1998). "Dimensionality reduction by random mapping: fast similarity computation for clustering". In: *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, pp. 413–418. DOI: 10.1109/IJCNN.1998.682302. URL: https://doi.org/10.1109/IJCNN.1998.682302 (visited on 02/01/2026).

Kaski, Samuel et al. (1998). "WEBSOM – Self-organizing maps of document collections." In: *Neurocomputing* 21.1, pp. 101–117. ISSN: 0925-2312. DOI: 10.1016/S0925-2312(98)00039-3. URL: https://doi.org/10.1016/S0925-2312(98)00039-3 (visited on 02/01/2026).

Leimeister, Chris Andre et al. (Mar. 2019). "Prot-SpaM: fast alignment-free phylogeny reconstruction based on whole-proteome sequences". In: *GigaScience* 8.3, pp. 1–14. ISSN: 2047217X. DOI: 10.1093/GIGASCIENCE/GIY148. URL: https://doi.org/10.1093/GIGASCIENCE/GIY148 (visited on 02/01/2026).

Lilleberg, Joseph, Yun Zhu, and Yanqing Zhang (Sept. 2015). "Support vector machines and Word2vec for text classification with semantic features". In: *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2015*, pp. 136–140. DOI: 10.1109/ICCI-CC.2015.7259377. URL: https://doi.org/10.1109/ICCI-CC.2015.7259377 (visited on 02/01/2026).

Ma, Long and Yanqing Zhang (Dec. 2015). "Using Word2Vec to process big text data". In: *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 2895–2897. DOI: 10.1109/BIGDATA.2015.7364114. URL: https://doi.org/10.1109/BIGDATA.2015.7364114 (visited on 02/01/2026).

Mikolov, Tomas et al. (Jan. 2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv*. DOI: 10.48550/arXiv.1301.3781. URL: https://doi.org/10.48550/arXiv.1301.3781 (visited on 02/01/2026).

Mitchell, J. and M. Lapata (Nov. 2010). "Composition in Distributional Models of Semantics". In: *Cognitive Science* 34.8, pp. 1388–1429. ISSN: 1551-6709. DOI: 10.1111/J.1551-6709.2010.01106.X. URL: https://doi.org/10.1111/J.1551-6709.2010.01106.X (visited on 02/01/2026).

Mueller, Andreas C (2026). *Wordcloud*. URL: https://github.com/amueller/word_cloud (visited on 02/01/2026).

OpenAI et al. (2023). "GPT-4 Technical Report". In: *arXiv*. DOI: 10.48550/arXiv.2303.08774. URL: https://doi.org/10.48550/arXiv.2303.08774 (visited on 02/01/2026).

Pedregosa, F et al. (2011). *Scikit-learn: Machine Learning in Python*. URL: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (visited on 02/01/2026).

Perico, Camila P et al. (2022). "Genomic landscape of the SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin". In: *Frontiers in Microbiology* 13. DOI: 10.3389/fmicb.2022.1037455. URL: https://doi.org/10.3389/fmicb.2022.1037455 (visited on 02/01/2026).

Radford, Alec and Karthik Narasimhan (2018). *Improving Language Understanding by Generative Pre-Training*. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (visited on 02/01/2026).

Raffel, Colin et al. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. DOI: 10.48550/ARXIV.1910.10683. URL: https://doi.org/10.48550/ARXIV.1910.10683 (visited on 02/01/2026).

Raittz, Roberto Tadeu et al. (2021). "Comparative genomics provides insights into the taxonomy of azoarcus and reveals separate origins of nif genes in the proposed azoarcus and aromatoleum genera". In: *Genes* 12, pp. 1–21. DOI: 10.3390/genes12010071. URL: https://doi.org/10.3390/genes12010071 (visited on 02/01/2026).

Robertson, Stephen (2004). "Understanding inverse document frequency: On theoretical arguments for IDF". In: *Journal of Documentation* 60.5, pp. 503–520. DOI: 10.1108/00220410410560582. URL: https://doi.org/10.1108/00220410410560582 (visited on 02/01/2026).

Rubel, Elisa Terumi et al. (Dec. 2016). "ProClaT, a new bioinformatics tool for in silico protein reclassification: case study of DraB, a protein coded from the draTGB operon in Azospirillum brasilense". In: *BMC bioinformatics* 17.Suppl 18. ISSN: 1471-2105. DOI: 10.1186/S12859-016-1338-5. URL: https://doi.org/10.1186/S12859-016-1338-5 (visited on 02/01/2026).

Scorzato, Luigi (Sept. 2024). "Reliability and Interpretability in Science and Deep Learning". In: *Minds and Machines* 34.3, pp. 1–31. ISSN: 15728641. DOI: 10.1007/S11023-024-09682-0. URL: https://doi.org/10.1007/S11023-024-09682-0 (visited on 02/01/2026).

Silva Filho, Antonio Camilo da et al. (2021). "Prediction and Analysis in silico of Genomic Islands in Aeromonas hydrophila". In: *Frontiers in Microbiology* 12. DOI: 10.3389/fmicb.2021.769380. URL: https://doi.org/10.3389/fmicb.2021.769380 (visited on 02/01/2026).

Tay, Yi et al. (Dec. 2022). "Efficient Transformers: A Survey". In: *ACM Computing Surveys* 55.6, pp. 1–28. ISSN: 1557-7341. DOI: 10.1145/3530811. URL: http://doi.org/10.1145/3530811 (visited on 02/01/2026).

Tshitoyan, Vahe et al. (July 2019). "Unsupervised word embeddings capture latent knowledge from materials science literature". In: *Nature* 571.7763, pp. 95–98. ISSN: 0028-0836. DOI: 10.1038/s41586-019-1335-8. URL: https://doi.org/10.1038/s41586-019-1335-8 (visited on 02/01/2026).

Van Rossum, Guido and Python Software Foundation (2026). *Python Language Reference*. URL: https://docs.python.org/3/reference (visited on 02/01/2026).

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: DOI: 10.48550/arXiv.1706.03762. URL: https://arXiv.org/abs/1706.03762 (visited on 02/01/2026).

Wang, Sinong et al. (2020). "Linformer: Self-Attention with Linear Complexity". In: DOI: 10.48550/arXiv.2006.04768. URL: http://doi.org/10.48550/arXiv.2006.04768 (visited on 02/01/2026).

Supplementary Material

Table S1: Methods in linguistic vectorization. An overview of concepts and methods in the field, presented in chronological order and accompanied by a brief description of their key contributions.

Year	Model/Technique	Key Contribution	References
1972	Inverse Document Frequency (IDF)	A statistical measure that emphasizes rare but informative terms, improving document retrieval and weighting in information retrieval systems.	Jones (1972)
1990	Recurrent Neural Networks (RNN)	A foundational architecture for sequential data processing, where hidden states retain temporal context, enabling early NLP sequence modeling.	Elman (1990)
1994	Random Indexing	A method for sparse semantic representations, inspired by cognitive principles, that reduces computational overhead.	Kanerva (1994) and Kanerva, Kristoferson, and Hols (2000)
1998	Random Mapping	A dimensionality reduction technique that preserves pairwise semantic similarities, facilitating efficient large-scale analysis.	Kaski (1998) and Samuel Kaski et al. (1998)
2003	Neural Language Model	A neural network-based approach that learns distributed word representations, capturing semantic relationships.	Bengio et al. (2003)
2013	Word2Vec	A word embedding approach that uses skip-gram and CBOW architectures to model semantic relationships via context windows.	Mikolov et al. (2013)
2016	FastText	Extends Word2Vec by incorporating subword information, enhancing representations for rare words and morphologically complex languages.	Bojanowski et al. (2016)
2018	GPT (Generative Pretrained Transformer)	A unidirectional Transformer model that excels in autoregressive text generation through large-scale pretraining.	Radford and Narasimhan (2018)
2018	BERT (Bidirectional Encoder Representations from Transformers)	A bidirectional Transformer model that achieves deep contextual understanding, setting new benchmarks across NLP tasks.	Devlin et al. (2018)
2019	T5 (Text-to-Text Transfer Transformer)	A unified text-to-text framework that casts all NLP tasks as generating text from text. Pre-trained via span corruption and fine-tuned with task instructions, it offers flexibility for generative and comprehension tasks.	Raffel et al. (2019)
2025	DeepSeek-R1	A model that combines multi-stage reinforcement learning and model distillation (based on Qwen/Llama) for efficient language alignment.	DeepSeek-AI et al. (2025)

Table S2: AMINOcode character substitution rules. Table showing the AMINOcode encoding rules. This case-insensitive method offers two approaches: reduced encoding (compact, limited to alphanumeric characters and basic symbols) and detailed encoding (comprehensive but longer).

Character	Detailed	Reduced	Character	Detailed	Reduced
a	YA	YA	x	W	W
b	E	E	z	A	A
c	C	C	w	YW	YW
d	D	D	y	YY	YY
e	YE	YE	0	YDA	YD
f	F	F	1	YDQ	
g	G	G	2	YDT	
h	H	H	3	YDH	
i	YI	YI	4	YDF	
j	I	I	5	YDI	
k	K	K	6	YDS	
l	L	L	7	YDE	
m	M	M	8	YDG	
n	N	N	9	YDN	
o	YQ	YQ	.	YPE	YP
p	P	P	,	YPC	
q	Q	Q	;	YPS	
r	R	R	!	YPW	
s	S	S	?	YPQ	
t	T	T	:	YPT	
u	YV	YV	space	YS	YS
v	V	V	exception	YK	YK

Table S3: Example of AMINOcode encoding and decoding.

Description	String
Original Text	SWeeP is amazing! 10 to 100 times faster than other techniques. #Bioinformatics #DiscoverSWeeP
Encoded (reduced)	SYWYEYEPYSYISYSYAMYAAYINGYPYSYDYDYSTYQYSYDYDYDYSTYIMYESYSFYASTYERYSTHYANYSYQTHYERYSTYECHNYIQYVYESYPYSYKEYIYQYINFYQRMYATYICSYSYKDYISCYQVYERSYWYEYEP
Encoded (detailed)	SYWYEYEPYSYISYSYAMYAAYINGYPWYSYDQYDAYSTYQYSYDQYDAYDAYSTYIMYESYSFYASTYERYSTHYANYSYQTHYERYSTYECHNYIQYVYESYPEYSYKEYIYQYINFYQRMYATYICSYSYKDYISCYQVYERSYWYEYEP
Decoded (reduced)	sweep is amazing. 99 to 999 times faster than other techniques. -bioinformatics -discoversweep
Decoded (detailed)	sweep is amazing! 10 to 100 times faster than other techniques. -bioinformatics -discoversweep

Table S4: Example of DNAbits encoding and decoding.

Description	String
Original Text	SWeeP is amazing! 10 to 100 times faster than other techniques. #Bioinformatics #DiscoverSWeeP
Encoded	TACCTCCCCCGCCCGCAACCAAGACGGCTATCAAGACAGCCTGCCAGCGGTCCGGCGTGCTCGCCAGAAAGACATAAATAAAGAACTCTTGCAAGACATAAATAAATAAAGAACTCCGGCCTGCCCGCTATCAAGAGCGCCAGCTATCACTCCCGCGATCAAGAACTCAGGCCAGCGTGCAAGATTGCACTCAGGCCCGCGATCAAGAACTCCCGCTAGCAGGCGTGCCGGCCATCCCTCCCGCTATCGTGAAAGATAGAGAACCGGCTTGCCGGCGTGCGCGCTTGCGATCCTGCCAGCACTCCGGCTAGCTATCAAGATAGAACACCGGCTATCTAGCTTGCGCTCCCGCGATCTACCTCCCCCGCCCGCAACC
Decoded	SWeeP is amazing! 10 to 100 times faster than other techniques. #Bioinformatics #DiscoverSWeeP

Table S5: Hardware specifications. Hardware specifications of the system used for analysis.

Component	Specifications
Processor (CPU)	Intel Core i5-3470 @ 3.20GHz (4 cores, 4 threads, 3.6GHz Turbo)
CPU Cache	L2: 1 MB, L3: 6 MB
Memory (RAM)	16 GB DDR3 @ 1333MHz (2 × 8GB Kingston 99U5471-060.A00LF modules)
RAM Configuration	Dual Channel (ChannelA-DIMM0, ChannelB-DIMM0)
Motherboard	Gigabyte H61M-S1 (Intel H61 chipset, LGA1155 socket)

Table S6: PubMed search specifications. PubMed search parameters and results used to construct the dataset for the thioredoxin usage example.

Description	Information
Search Date	12 April 2025
Interface	Entrez Direct command line tool
Search String	`thioredoxin AND english[language] AND hasabstract NOT "Published Erratum"[Publication Type]`
Output File	thioredoxin_2025.04.12.medline
Command Used	`esearch -db pubmed -query "$SEARCH_STRING" \| efetch -format medline > "$OUTPUT_FILE"`
Result	14,984 documents

Table S7: Documents related to "leukemia" and "pregnancy". Titles and PubMed IDs (PMID) of documents most closely associated with the target words "leukemia" and "pregnancy".

Target Word	Title	PubMed ID
leukemia	Possible roles of an adult T-cell leukemia (ATL)-derived factor/thioredoxin in the drug resistance of ATL to adriamycin.	9116292
	A novel thioredoxin reductase inhibitor inhibits cell growth and induces apoptosis in HL-60 and K562 cells.	18196608
	Inhibition of thioredoxin reductase by auranofin induces apoptosis in adriamycin-resistant human K562 chronic myeloid leukemia cells.	21699084
	Ethaselen: a novel organoselenium anticancer agent targeting thioredoxin reductase 1 reverses cisplatin resistance in drug-resistant K562 cells by inducing apoptosis.	28471109
	Inhibition of the Nrf2-TrxR Axis Sensitizes the Drug-Resistant Chronic Myelogenous Leukemia Cell Line K562/G01 to Imatinib Treatments.	31828114
	Synergism between thioredoxin reductase inhibitor ethaselen and sodium selenite in inhibiting proliferation and inducing death of human non-small cell lung cancer cells.	28757135
	Shikonin inhibits gefitinib-resistant non-small cell lung cancer by inhibiting TrxR and activating the EGFR proteasomal degradation pathway.	27864022
	A thioredoxin reductase inhibitor induces growth inhibition and apoptosis in five cultured human carcinoma cell lines.	15982805
	Thioredoxin-1 inhibitor PX-12 induces human acute myeloid leukemia cell apoptosis and enhances the sensitivity of cells to arsenic trioxide.	25197347
	CD40 ligation inhibits IL-2 and SAC+IL-2 induced proliferation in chronic lymphocytic leukaemia cells.	9201312
pregnancy	Changes in maternal serum thioredoxin (TRX) levels after delivery in preeclamptic and normotensive pregnant women.	21250889
	Beneficial effects of dietary fibre supplementation of a high-fat diet on fetal development in rats.	21486515
	Effect of different iodide intake during pregnancy and lactation on thyroid and cardiovascular function in maternal and offspring rats.	37506535
	Serum manganese superoxide dismutase and thioredoxin are potential prognostic markers for hepatitis C virus-related hepatocellular carcinoma.	22171130
	Elevation of serum thioredoxin levels in patients with type 2 diabetes.	11972307
	Patients with Osteoarthritis and Kashin-Beck Disease Display Distinct CpG Methylation Profiles in the DIO2, GPX3, and TXRND1 Promoter Regions.	33455417
	Measurable serum markers of oxidative stress response in women with endometriosis.	18206876
	Selenium inadequacy hampers thyroid response of young children after iodine repletion.	30262294
	Elevation of blood thioredoxin in hemodialysis patients with hepatitis C virus infection.	12753316
	Aortic Intima-Media Thickness is Increased in Neonates of Mothers with Gestational Diabetes Mellitus: The Role of Thioredoxin-Interacting Protein as a Marker of Oxidative Stress.	37518994

Table S8: Top 30 most frequent representative terms for document clusters. The columns to the right indicate the frequency of each term as the most representative (highest TF-IDF score) for a document within its cluster. Abbreviation: Freq. = Frequence.

#	Cluster 0	Freq.	Cluster 1	Freq.	Cluster 2	Freq.
1	pdi	64	txnip	122	trx	76
2	trx	52	nrf2	44	se	72
3	cys	51	ask1	41	selenium	71
4	ntrc	42	nlrp3	41	trxr	71
5	disulfide	39	se	36	2	55
6	prx	39	0	29	gold	55
7	dsba	38	trx	28	redox	49
8	fusion	37	trx1	26	trxr1	41
9	peptide	29	endothelial	23	genes	36
10	sec	27	glucose	22	gsh	35
11	nadp	25	txndc5	22	ros	35
12	trxa	25	hcc	21	complexes	35
13	proteins	24	diabetic	21	prx	30
14	thioredoxins	22	kappab	19	mitochondrial	28
15	arsenate	22	exercise	19	trx1	28
16	tuberculosis	22	trx2	19	txnip	27
17	dna	21	trxr1	19	auranofin	26
18	glutaredoxin	21	ad	18	thyroid	25
19	chloroplast	21	resveratrol	18	tr	24
20	ftr	21	hg	18	selenoproteins	24
21	disulphide	20	px	17	cancer	23
22	tgr	20	txnrd1	17	nitrosylation	21
23	t7	19	af	17	adf	19
24	folding	19	5p	17	b	19
25	methionine	19	hif	16	selenite	19
26	er	18	mice	16	drought	19
27	fbpase	18	hypoxia	16	txnrd1	19
28	domain	18	mi	16	h2o2	18
29	trxs	18	mir	16	p53	18
30	grx	17	diabetes	15	compounds	18

Table S9: Analysis of document clusters.

Cluster	Title	Description
0	Bacterial Redox and Biotechnological Applications	This cluster centers on the fundamental and applied study of microbial thioredoxin systems, primarily in bacterial models. Research focuses on elucidating the structure and function of redox enzymes involved in processes like protein folding and disulfide bond formation. A strong biotechnological theme is evident, with a focus on metabolic engineering, enzyme kinetics, and understanding microbial adaptation to environmental stresses.
1	Oxidative Stress in Metabolic and Degenerative Diseases	Research in this cluster investigates the role of thioredoxin-related pathways in the pathophysiology of chronic mammalian diseases. The central theme is redox imbalance and oxidative stress as drivers of tissue dysfunction. Studies explore disease mechanisms and therapeutic interventions in contexts such as liver disease, cardiac dysfunction, diabetes, neurodegeneration, and aging.
2	Eukaryotic Redox Biology and Developmental Systems	This cluster encompasses studies on thioredoxin networks in eukaryotic organisms, emphasizing developmental biology, reproduction, and organismal stress adaptation. A key focus is on redox regulation of fertility, embryogenesis, and developmental signaling. Research frequently utilizes genetic animal models to dissect selenium-dependent and independent pathways.

Table S10: Examples of documents from each cluster. Documents selected using different approaches – centroid-proximate and random sampling – with interpretations for cluster membership. Abbreviations: Cl = Cluster; S = Selection method; c = centroid-proximate; r = random.

Title	Cl	S	Interpretation for Cluster Membership
Elucidating paramylon and other carbohydrate metabolism in *Euglena gracilis*: Kinetic characterization, structure and cellular localization of UDP-glucose pyrophosphorylase.	0	c	Focus on enzymatic characterization and carbohydrate metabolism, aligning with Cluster 0's emphasis on metabolic pathways and biotechnological applications.
Sulfolobus solfataricus thiol redox puzzle: characterization of an atypical protein disulfide oxidoreductase.	0	c	Study of disulfide-related enzymes in prokaryotes fits Cluster 0's focus on microbial redox systems and protein folding mechanisms.
Thioredoxin system in obligate anaerobe *Desulfovibrio desulfuricans*: Identification and characterization of a novel thioredoxin 2.	0	c	Investigation of redox systems in anaerobic bacteria matches Cluster 0's interest in microbial adaptations and enzyme functions.
Relevance of FXR-p62/SQSTM1 pathway for survival and protection of mouse hepatocytes and liver, especially with steatosis.	1	c	Research on liver disease and metabolic stress aligns with Cluster 1's focus on oxidative stress in metabolic disorders.
Inhibitory effect of alpha-lipoic acid on thioacetamide-induced tumor promotion through suppression of inflammatory cell responses in a two-stage hepatocarcinogenesis model in rats.	1	c	Study of oxidative stress and inflammation in liver disease fits Cluster 1's emphasis on redox imbalance in chronic diseases.
Coptis chinensis Franch. exhibits neuroprotective properties against oxidative stress in human neuroblastoma cells.	1	c	Investigating oxidative stress in neurodegenerative contexts matches Cluster 1's focus on redox-related diseases.
Seaweed extracts and unsaturated fatty acid constituents from the green alga *Ulva lactuca* as activators of the cytoprotective Nrf2-ARE pathway.	2	c	Study of redox-regulated cytoprotective mechanisms aligns with Cluster 2's interest in eukaryotic stress responses.
Improved tag-switch method reveals that thioredoxin acts as depersulfidase and controls the intracellular levels of protein persulfidation.	2	c	Research on redox regulation in cellular systems aligns with Cluster 2's focus on eukaryotic redox biology.
Epididymal specific, selenium-independent GPX5 protects cells from oxidative stress-induced lipid peroxidation and DNA mutation.	2	c	Focus on selenoprotein-independent redox defense in development aligns with Cluster 2's emphasis on eukaryotic redox systems.
Vitamin K epoxide reductase prefers ER membrane-anchored thioredoxin-like redox partners.	0	r	Study of redox enzyme interactions aligns with Cluster 0's biotechnological and enzymatic focus.
Cloning, characterization, and expression analysis of a thioredoxin from orange-spotted grouper (*Epinephelus coioides*).	0	r	Characterization of redox proteins in a biological system fits Cluster 0's emphasis on enzyme function and metabolic studies.
Purification and properties of methyl sulfoxide reductases from rat kidney.	0	r	Investigation of reductases applications aligns with Cluster 0's biotechnological approach.
Long-term administration of low-dose selenium nanoparticles with different sizes aggravated atherosclerotic lesions and exhibited toxicity in apolipoprotein E-deficient mice.	1	r	Study of oxidative stress and toxicity in a disease model matches Cluster 1's focus on redox imbalance in metabolic disorders.
Iridium (III) complexes induce cervical carcinoma apoptosis via disturbing cellular redox homeostasis disorder and inhibiting PI3K/AKT/mTOR pathway.	1	r	Research on redox disruption in cancer aligns with Cluster 1's interest in oxidative stress and disease mechanisms.
Gene expression profiling in INS-1 cells overexpressing thioredoxin-interacting protein.	1	r	Investigation of redox-related gene regulation fits Cluster 1's focus on metabolic and degenerative diseases.

Continued on the next page

Table S10 – Continuation of the previous page

Integrative Model of Oxidative Stress Adaptation in the Fungal Pathogen *Candida albicans*.	2	r	The study's focus on oxidative stress adaptation in a eukaryotic pathogen directly aligns with Cluster 2's themes of redox-regulated signaling and resilience in eukaryotic systems.
Transcriptomic and Proteomic Analysis of *Oenococcus oeni* Adaptation to Wine Stress Conditions.	2	r	Although the target organism is prokaryotic, its relation to wine alcoholic fermentation – a process strongly associated with eukaryotes – may have created semantic affinity with the terms and themes of Cluster 2.
Definition of receptor binding sites on human interleukin-11 by molecular modeling-guided mutagenesis.	2	r	Focus on human signaling mechanisms aligns with Cluster 2's relationship to eukaryotic regulatory pathways.

Figure S1: Elbow method for optimal cluster selection. (A) Variance explained versus number of clusters (k) for word embeddings, with the chosen $k = 7$. (B) Equivalent analysis for document embeddings, selecting $k = 3$. Dashed lines indicate the elbow criteria.

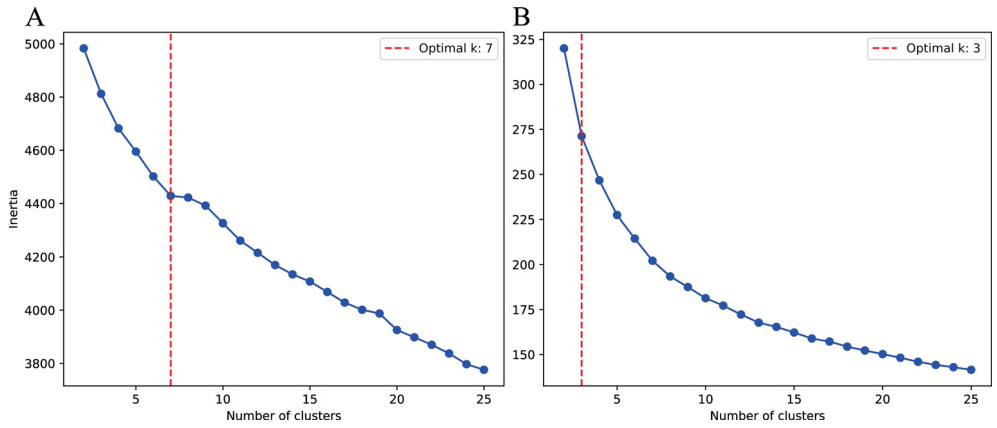


Figure S2: Word clouds of document clusters. Word clouds depicting the most frequent terms in each document cluster. Word size corresponds to relative frequency within each grouping.

Box S1: Usage example: demonstration. This box illustrates the conceptual workflow described in the main text using a simplified, artificial corpus of biomedical sentences. All vectors and computations shown are symbolic placeholders for demonstration purposes only and do not correspond to actual computational outputs. The example demonstrates the pipeline's logical progression.

Step 1: Data Collection and Preparation
A PubMed search for "thioredoxin" yields a corpus of 7 representative abstracts:
0. Document 0: "thioredoxin regulates redox signaling in cancer cells"
1. Document 1: "redox balance and oxidative stress in cardiovascular disease"
2. Document 2: "thioredoxin interacts with apoptosis signaling pathways"
3. Document 3: "oxidative stress markers in Alzheimer's disease"
4. Document 4: "the role of thioredoxin in diabetes-related oxidative stress"
5. Document 5: "signaling pathways in neurodegenerative diseases and oxidative damage"
6. Document 6: "redox signaling mediates both cellular protection and pathology"

Step 2: Text Preprocessing
- Convert to lowercase and tokenize.
- Remove stopwords ("the", "in", "and", etc.).
- Apply TF-IDF filtering (threshold ≥ 0.5) to select domain-relevant terms.

Step 3: SWeePtex Embedding Generation
Preliminary Document Vectors: Convert each document to BSL format using AMINOcode, then project to SWeePtex vectors $d_i^{(0)} \in \mathbb{R}^{1,200}$.
Word Embeddings: Compute $e_w = \frac{1}{|\mathcal{D}(w)|} \sum_{D_i \in \mathcal{D}(w)} d_i^{(0)}$ for each term.
Refined Document Embeddings: Compute $d_i = \frac{1}{L_i} \sum_{j=1}^{L_i} e_{w_{ij}}$ for each document.
Result: 7 document embeddings $d_0, \ldots, d_6$ and 12 word embeddings for the filtered vocabulary.

Step 4: Dimensionality Reduction
Apply PCA to reduce word and document embeddings from 1,200 to 50 dimensions.

Step 5: Word-Level Analysis
Optimal Cluster Determination: Elbow method on word embeddings suggests $k_{\text{words}} = 2$.
k-means Clustering: Words group into:
- Cluster 0: thioredoxin, redox, signaling, pathways, mediates, cellular
- Cluster 1: cancer, cardiovascular, alzheimer, oxidative, stress, diabetes, neurodegenerative, pathology

Step 6: Document-Level Analysis
TF-IDF Salient Terms: For each document, extract top terms:
- Document 0: thioredoxin, redox, signaling, cancer
- Document 1: redox, oxidative, stress, cardiovascular
- Document 2: thioredoxin, apoptosis, signaling, pathways
- Document 3: oxidative, stress, alzheimer, markers
- Document 4: thioredoxin, diabetes, oxidative, stress
- Document 5: signaling, pathways, neurodegenerative, oxidative, damage
- Document 6: redox, signaling, mediates, cellular, protection, pathology

Document Clustering: Elbow method suggests $k_{\text{docs}} = 2$:
- Cluster 0: Documents 0, 2, 4, 6 (thioredoxin/redox signaling focus)
- Cluster 1: Documents 1, 3, 5 (oxidative stress/disease focus)

Continued on the next page

Box S1 – Continuation of the previous page

Step 7: Cluster Representation
Centroid-Proximate Selection (most representative):
- Cluster 0: Document 2 (thioredoxin-apoptosis signaling)
- Cluster 1: Document 1 (oxidative stress in cardiovascular disease)

Random Sampling Selection (variability):
- Cluster 0: Document 6 (redox signaling in cellular protection/pathology)
- Cluster 1: Document 5 (signaling in neurodegenerative oxidative damage)

Step 8: Output
The pipeline generates visualizations, word clouds, and tabular summaries.

Step 9: Human interpretation
In analysis, word clustering separates mechanistic terms (e.g., thioredoxin, redox, signaling) from pathological/disease terms (e.g., oxidative stress, alzheimer, diabetes). Document clustering also represents these themes; Cluster 0 focuses on thioredoxin and redox signaling mechanisms, while Cluster 1 centers on oxidative stress across different disease contexts. Documents near the centroids provide representative examples of each core theme, while random sampling reveals thematic variations and mixed-concept documents (e.g., Document 6 bridges signaling and pathology).

Box S2: SWeePtex embedding process: demonstration. This box illustrates the SWeePtex embedding procedure described in the main text, using a simplified corpus of three biomedical sentences. The example is purely illustrative. It demonstrates the bidirectional embedding process: from preliminary document vectors to word embeddings, and then to refined document embeddings.

Corpus of 3 Abstracts:
0. Document 0: "cell redox signaling"
1. Document 1: "redox cell stress"
2. Document 2: "stress signaling pathway"

Vocabulary:
$\mathcal{W} = \{\text{cell, redox, signaling, stress, pathway}\}$

Step 1: Preliminary Document Vectors
Each document D_i is converted to its preliminary SWeePtex vector $d_i^{(0)} \in \mathbb{R}^m$:
$d_0^{(0)}$ = vector for "cell redox signaling"
$d_1^{(0)}$ = vector for "redox cell stress"
$d_2^{(0)}$ = vector for "stress signaling pathway"

Step 2: Word Embeddings
Each word embedding e_w is computed as the average of preliminary vectors from documents containing that word:
$e_{\text{cell}} = \frac{1}{2}(d_0^{(0)} + d_1^{(0)})$ (appears in Docs 0 and 1)

$e_{\text{redox}} = \frac{1}{2}(d_0^{(0)} + d_1^{(0)})$ (appears in Docs 0 and 1)

$e_{\text{signaling}} = \frac{1}{2}(d_0^{(0)} + d_2^{(0)})$ (appears in Docs 0 and 2)

$e_{\text{stress}} = \frac{1}{2}(d_1^{(0)} + d_2^{(0)})$ (appears in Docs 1 and 2)

$e_{\text{pathway}} = d_2^{(0)}$ (appears only in Doc 2)

Step 3: Refined Document Embeddings
Each refined document embedding d_i is computed as the average of its word embeddings:
$d_0 = \frac{1}{3}(e_{\text{cell}} + e_{\text{redox}} + e_{\text{signaling}})$
(contains: cell, redox, signaling)
$d_1 = \frac{1}{3}(e_{\text{redox}} + e_{\text{cell}} + e_{\text{stress}})$
(contains: redox, cell, stress)
$d_2 = \frac{1}{3}(e_{\text{stress}} + e_{\text{signaling}} + e_{\text{pathway}})$
(contains: stress, signaling, pathway)

Bidirectional Context Propagation:
1. Documents 0 and 1 share words "cell" and "redox", so their refined embeddings become more similar.
2. Document 2 shares "signaling" with Document 0 and "stress" with Document 1, creating indirect relationships.
3. The word "pathway" in Document 2 receives context only from that document, making it a distinctive feature.

Result: The final embeddings d_0, d_1, d_2 reflect not only the original document content, but also the shared contextual information across the corpus. For instance, d_0 and d_1 become more similar because they share the words "cell" and "redox", whose embeddings are themselves informed by both documents.

Supplementary Material References

Bengio, Yoshua et al. (2003). "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3, pp. 1137–1155. DOI: 10.5555/944919.944966. URL: https://dl.acm.org/doi/abs/10.5555/944919.944966 (visited on 02/01/2026).

Bojanowski, Piotr et al. (2016). *Enriching Word Vectors with Subword Information*. DOI: 10.48550/arXiv.1607.04606. URL: https://doi.org/10.48550/arXiv.1607.04606 (visited on 02/01/2026).

DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. DOI: 10.48550/arXiv.2501.12948. URL: https://doi.org/10.48550/arXiv.2501.12948 (visited on 02/01/2026).

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv*. DOI: 10.48550/arXiv.1810.04805. URL: https://doi.org/10.48550/arXiv.1810.04805 (visited on 02/01/2026).

Elman, Jeffrey L (1990). "Finding Structure in Time". In: *Cognitive Science* 14.2, pp. 179–211. DOI: 10.1207/s15516709cog1402_1. URL: https://doi.org/10.1207/s15516709cog1402_1 (visited on 02/01/2026).

Jones, Sparck K (1972). *A Statistical Interpretation of Term Specificity and its Application in Retrieval*. URL: https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf (visited on 02/01/2026).

Kanerva, Pentti (1994). "The Spatter Code for Encoding Concepts at Many Levels". In: *ICANN '94*. London: Springer London, pp. 226–229. DOI: 10.1007/978-1-4471-2097-1_52. URL: https://doi.org/10.1007/978-1-4471-2097-1_52 (visited on 02/01/2026).

Kanerva, Pentti, Jan Kristoferson, and Anders Hols (2000). *Random Indexing of Text Samples for Latent Semantic Analysis*. URL: https://escholarship.org/uc/item/5644k0w6 (visited on 02/01/2026).

Kaski, S (1998). "Dimensionality reduction by random mapping: fast similarity computation for clustering". In: *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, pp. 413–418. DOI: 10.1109/IJCNN.1998.682302. URL: https://doi.org/10.1109/IJCNN.1998.682302 (visited on 02/01/2026).

Kaski, Samuel et al. (1998). "WEBSOM – Self-organizing maps of document collections." In: *Neurocomputing* 21.1, pp. 101–117. ISSN: 0925-2312. DOI: 10.1016/S0925-2312(98)00039-3. URL: https://doi.org/10.1016/S0925-2312(98)00039-3 (visited on 02/01/2026).

Mikolov, Tomas et al. (Jan. 2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv*. DOI: 10.48550/arXiv.1301.3781. URL: https://doi.org/10.48550/arXiv.1301.3781 (visited on 02/01/2026).

Radford, Alec and Karthik Narasimhan (2018). *Improving Language Understanding by Generative Pre-Training*. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (visited on 02/01/2026).

Raffel, Colin et al. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. DOI: 10.48550/ARXIV.1910.10683. URL: https://doi.org/10.48550/ARXIV.1910.10683 (visited on 02/01/2026).

2.2 ARTIGO 2 – SWEEPTEX-EMB: AVALIAÇÃO QUANTITATIVA

O Artigo 2 apresenta uma avaliação quantitativa do método SWeePtex-Emb para a geração de incorporações vetoriais de texto. O estudo conduz testes no *Massive Text Embedding Benchmark* (MTEB), utilizando modelos treinados no corpus DBpedia-14, que contém 630.000 documentos.

O SWeePtex-Emb atinge a acurácia de 0,80 na tarefa de classificação do DBpedia. Esse resultado posiciona seu desempenho abaixo das arquiteturas mais bem posicionadas no *benchmark*, porém superior ao de alguns modelos compactos.

A análise de correlação mostra que as pontuações do SWeePtex-Emb para as tarefas testadas apresentam alta correlação (coeficientes entre 0,72 e 0,93) com as de diversos modelos neurais avaliados no MTEB. Esse resultado sugere que, apesar das diferenças fundamentais em seus mecanismos de geração, as representações do SWeePtex-Emb capturam padrões nos dados que se alinham aos identificados por abordagens de aprendizagem profunda.

A avaliação também identifica fragilidades específicas. O método apresenta desempenho inferior em tarefas que exigem a compreensão de relações semânticas contextuais complexas, como a similaridade textual e a classificação em pares. Essa observação indica uma capacidade reduzida para modelar a sinonímia contextual e as relações de correferência não explícitas, possivelmente relacionadas ao processo de construção semântica baseado em médias vetoriais.

A discussão aborda desafios experimentais inerentes à avaliação comparativa. Um deles é a dificuldade em quantificar técnicas de modelagem em domínios altamente especializados, em que a falta de dados rotulados impede uma análise quantitativa robusta. Este ponto é exemplificado pelo estudo no domínio da Yoga (Ieger-Raittz et al., 2025), que, ao empregar o método SWeePtex para exploração da literatura, representa uma validação qualitativa relevante, mas não pode avançar para uma comparação numérica direta devido à ausência de um *dataset* supervisionado específico.

Outra limitação discutida é a dificuldade em comparar de forma justa com LLMs de alto desempenho. Tais modelos, especialmente os mais bem posicionados no MTEB, são treinados com volumes massivos de dados, frequentemente indisponíveis ou irreproduzíveis, o que inviabiliza a replicação equitativa das condições de treinamento. Embora uma abordagem potencial para contornar esse problema seja treinar versões de arquiteturas semelhantes a partir de um conjunto de dados controlado, esse processo não é trivial, o que acrescenta uma camada adicional de complexidade metodológica à avaliação comparativa.

Em conclusão, o Artigo 2 posiciona o SWeePtex-Emb como uma alternativa às

técnicas de aprendizagem profunda. Os resultados demonstram que ele apresenta desempenho comparável ao de modelos destilados de pequeno porte. Ao mesmo tempo, as altas correlações observadas indicam que seus padrões de representação se assemelham até mesmo aos de modelos de maior escala. A ausência de um platô de desempenho em função do volume de dados sugere ainda um potencial para ganhos adicionais em estudos futuros.

SWeePtex-Emb: Benchmarking Random Projection-Based Text Embeddings Inspired by Bioinformatics

Diogo de J. S. Machado[1], Leonardo Vicenzi[1], Fábio de O. Pedrosa[1], and Roberto T. Raittz[1]

[1]Federal University of Paraná (UFPR)

Abstract

SWeePtex-Emb is a text embedding methodology that bridges natural language processing and Bioinformatics by applying structured random projections originally conceived for biological sequence analysis to natural language. The approach transforms textual input into Biological Sequence-Like (BSL) representations using the SWeeP algorithm to generate task-adaptive embeddings directly from raw text. This offers a practical alternative for constructing domain-specific embedding models from scratch. This study trains models on the DBpedia14 dataset, comprising 630,000 documents, and evaluates them on the Massive Text Embedding Benchmark (MTEB), focusing primarily on DBpediaClassification while also performing exploratory evaluations across other tasks. The effect of data scale is analyzed using progressively larger training subsets. Results show that, when trained on the full dataset, SWeePtex-Emb achieves 0.80 accuracy on DBpediaClassification and demonstrates at least 0.72 correlation with general-purpose embedding models across diverse MTEB tasks. This high correlation indicates that the method's embeddings effectively capture fundamental semantic structures, aligning closely with patterns learned by large-scale neural models despite its substantially simpler and more efficient architecture. Critically, the full training process executes in under 3 hours on standard desktop hardware. This practical runtime, combined with the observed performance scaling with dataset size, suggests the method is viable for training on even larger corpora to further reinforce the embedding of coherent semantic patterns. Future projects may refine the projection mechanism, enhance contextual embedding strategies, investigate hybrid architectures, test alternative training corpora, and explore novel applications.

Keywords: Text embedding. Bioinformatics-inspired. Benchmark.

1 Introduction

The rising complexity of state-of-the-art text embedding models, typically implemented as Large Language Models (LLMs) with deep neural architectures trained on massive general-purpose datasets, presents significant challenges. Although they deliver high performance, these models require substantial computational resources, raising sustainability concerns (Schwartz et al. 2020). Currently, scientific research is increasingly favoring running LLMs locally on standard workstations rather than relying on cloud services. This paradigm shift, facilitated by efficient open source models, offers crucial advantages, including enhanced privacy for sensitive data, reduced operational costs, improved reproducibility, and greater customization flexibility (Hutson 2024). Although the prevailing trend favors fine-tuning pre-trained generic models for specific applications, specialized domains may benefit from training models from scratch to mitigate biases inherent in general-purpose training data (Li, Hu, and Wang 2025; Yao et al. 2022).

The SWeeP (Spaced Words Projection) algorithm represents one such alternative, employing random projection techniques initially developed for biological sequence analysis (Silva Filho et al. 2021; De Pierri et al. 2020; Perico et al. 2022; Raittz et al. 2021). The SWeePtex adaptation extends this methodology to natural language processing by transformation of textual data into Biological Sequence-Like (BSL) representations using the Biotext framework (Machado et al. 2021), preserving the computational advantages of the original method while addressing the requirements of linguistic analysis. A previous study by our group on Yoga (Ieger-Raittz et al. 2025) demonstrated the qualitative applicability of this approach through visual exploration, which integrated SWeeP-generated vector representations within an NLP framework for text analysis.

Here, we demonstrate the training and quantitative evaluation of SWeePtex-Emb (SWeePtex Embedding) models. Training is performed on the DBpedia dataset (Lehmann et al. 2015), specifically DBpedia-14 (X. Zhang, Zhao, and LeCun 2016), which contains 630,000 entries. DBpedia is a multilingual dataset of encyclopedic knowledge derived from Wikipedia, while DBpedia-14 is a subset containing primarily English data. The evaluation is carried out using the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al. 2022), focusing on the DBpediaClassification task, but also includes other functions of different types and domains for exploratory purposes.

This study aims to contribute to ongoing discussions about sustainable NLP development by quantifying the performance characteristics of an efficient alternative to conventional embedding approaches, building upon our previous qualitative demonstration of the method's applicability in specialized domains. Furthermore, the study promotes the theoretical understanding of the composition of linguistic vectors.

2 Method

This study systematically evaluates text embedding models using the SWeeP algorithm on the DBpedia-14 corpus, employing a carefully designed experimental pipeline on a personal desktop computer. The methodology begins with comprehensive data preparation, processing documents as concatenated title-content pairs from the DBpedia-14 dataset, which includes 560,000 training documents and 70,000 test documents.

Two distinct model configurations are evaluated: a train-only model using exclusively the 560,000 training documents (sweeptex-emb-dbpedia-train-only) and a full corpus model trained on the combined training and test sets, totaling 630,000 documents (sweeptex-emb-dbpedia-full). During embedding generation, a minimum word frequency threshold of 3 occurrences is enforced, while the vocabulary remains unrestricted in size.

The evaluation assesses model performance across available MTEB English tasks using the HuggingFace leaderboard (accessed 14 August 2025) as the primary reference. DBpediaClassification serves as the primary ranking metric due to its direct relevance to the model's training domain and its inclusion in the main benchmark ranking. InstructionRetrieval tasks are excluded because they require textual instructions, which are incompatible with our embedding-only approach. TempReasonL1 tasks are excluded because they require temporal reasoning beyond pure semantic representation. Additionally, tasks exceeding our computational constraints are not considered.

The evaluation covers tasks across four MTEB categories: Classification (e.g., DBpediaClassification, PoemSentimentClassification, ToxicConversationsClassification), Clustering (e.g., ArXivHierarchicalClusteringP2P, ArXivHierarchicalClusteringS2S, BigPatentClustering), Retrieval (e.g., AILAStatutes, ArguAna, HagridRetrieval), and Semantic Textual Similarity (e.g., SICK-R, STS12, STS13).

To provide a comprehensive context, the analysis compares our best-performing SWeePtex-Emb model with three distinct model groups. First, the top three overall performers on the leaderboard establish a performance ceiling. Second, the three models that follow ours provide a direct rank-based comparison. Third, the three most parameter-efficient models that still outperform ours are selected to highlight efficiency trade-offs. All compared models must have their parameter counts available on the MTEB leaderboard.

To assess performance, the study employs a dual strategy evaluation framework. The first approach fixes the embedding dimensionality at 1,200 and progressively increases the training data size from 40,000 to 560,000 documents in increments of 40,000 to analyze the impact of dataset scale. The second approach maintains the sample size at 560,000 documents while systematically varying the number of PCA components from 100 to 1,200 in increments of 100 to isolate the effects of dimensionality reduction. Both strategies are benchmarked using the MTEB DBpe-

diaClassification task, with computational efficiency monitored throughout to ensure completion within the system's capabilities.

3 Result

The complete research material for creating, evaluating, and analyzing SWeePtex-Emb models is available in the Zenodo repository[1]. The content includes all source code, configuration files, experimental results, and visualization outputs.

The embedding generation pipeline requires almost 3 hours (10,545.61 seconds) of total execution time for the full corpus model. The computationally intensive phase, which consists of executing the SWeeP algorithm on the training and test sets, consumes 99.53% of this period (10,496.01 seconds; 9,127.30 for the training set and 1,368.71 for the test set).

Table 1 shows the models selected for comparison based on the criteria cited in the methodology, together with the corresponding parameter numbers, and the result of the DBpediaClassification. The result file set, with all tasks and models, is available in the experimental material. Figure 1 shows the correlation between the models based on the metric vectors.

Configuration experiments demonstrate performance scaling. Figure 2A tracks the impact of PCA dimensionality, revealing a performance plateau beyond 800 components despite a transient drop at 700 components. Figure 2B plots accuracy against the size of the training corpus, showing incremental improvements with occasional regressions at specific data volumes.

Table 1: Benchmark model selection and performance comparison. The table presents nine models selected through systematic evaluation against the SWeePtex-Emb, organized by selection criteria, along with their performance scores on the DBpediaClassification task (Accuracy metric). Models are grouped into three categories: (1) Top performers (highest absolute scores), (2) Compact models outperforming our solution, and (3) Comparable alternatives slightly below our performance threshold. Parameter counts reflect model complexity.

Model	Parameters	Selection	Rank	Accuracy
sweeptex-emb-dbpedia-full	-	SWeePtex-Emb	-	0.809619
sweeptex-emb-dbpedia-train-only	-		-	0.801270
Qwen3-Embedding-8B	7B	Top performer	1st	0.992578
Qwen3-Embedding-4B	4B		2nd	0.992578
Qwen3-Embedding-0.6B	595M		3rd	0.988428
potion-base-8M	7M	Small model	1st	0.819189
Squirtle	15M		2nd	0.835449
Venusaur	15M		3rd	0.907129
M2V_base_output	7M	Model below	1st	0.809570
rubert-base-cased	1B		2nd	0.804736
rubert-tiny-turbo	29M		3rd	0.796484

[1] https://doi.org/10.5281/zenodo.18370869

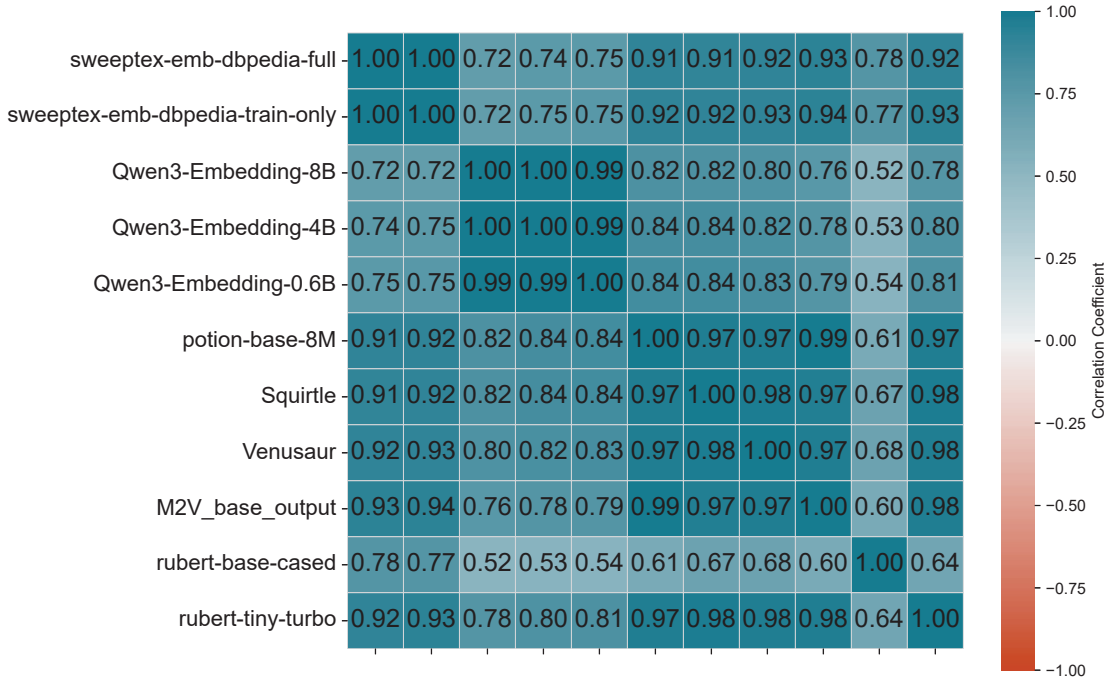


Figure 1: Correlation heatmap of embedding model performance metrics. Pearson correlation coefficients (ranging from 0.82 to 0.95) between SWeePtex-Emb variants and nine comparator models across MTEB tasks.

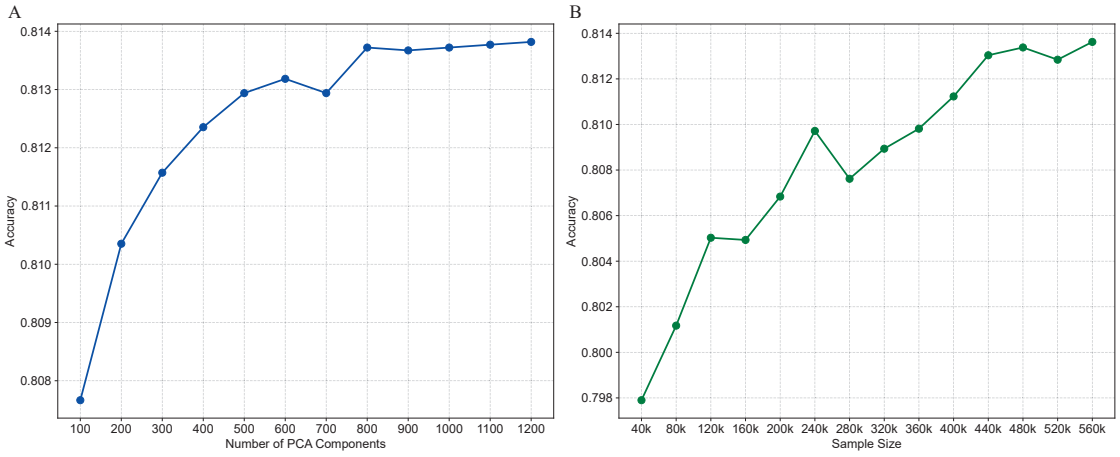


Figure 2: Performance scaling with training corpus size and PCA dimensionality. (A) Accuracy as PCA components vary from 100 to 1,200 (fixed 560K training documents). (B) Accuracy on DBpediaClassification as training corpus size increases from 40K to 560K documents (fixed 1,200 PCA dimensions).

4 Discussion

The results show that our approach, based on DBpediaClassification ordination, achieves an accuracy of 0.8096. While this performance is below that of the top models on the MTEB leaderboard, it is close to that of some compact models (Table 1). Analyzing the complete results table in the experimental material, it is notable that a task with a domain shared with the training set (encyclopaedic) also

shows a reasonable outcome: HagridRetrieval with an nDCG@10 of 0.7206.

The performance on the HagridRetrieval task is significant because this task shares the same encyclopaedic domain as the DBpedia dataset used for training. This alignment of the domain indicates that the SWeePtex-Emb model effectively captures semantic relationships within a consistent knowledge base. The model's ability to retrieve relevant documents within this shared domain demonstrates its practical utility. This domain-specific strength complements the more general classification performance shown in DBpediaClassification, providing a comprehensive view of the model's capabilities across different NLP tasks. Furthermore, performance on a retrieval task is particularly relevant for applications such as HTML-TM, which relies on effective retrieval to identify semantically related texts and terms.

The model with the fewest parameters that outperforms the SWeePtex-Emb result in DBpediaClassification is the potion-base-8M, a distilled model with 7M parameters. Among the three selected models with performance slightly below sweeptex-emb-dbpedia-full in the DBpedia Classification task, the best-performing is M2V_base_output, also a distilled model. We consider model distillation to be a projection of one model onto another, akin to a teacher teaching a student, as presented by DeepSeek-R1 (DeepSeek-AI et al. 2025). Thus, it is fitting that the random projection yielded similar results to those in this category. However, it is essential to emphasize that these distilled models depend on the prior existence of large "teacher" models, which are computationally expensive to train. Thus, SWeePtex-Emb presents an alternative that does not have this dependence.

Interestingly, the two other models that are just below SWeePtex-Emb performance are variations of RuBERT (Kuratov and Arkhipov 2019), which specializes in Russian. Their proximity to SWeePtex-Emb's performance may stem from RuBERT's training corpus, which consists mainly of Wikipedia data, an encyclopedic source like DBpedia.

SWeePtex-Emb's metrics show a high positive correlation with those of other models, ranging from 0.72 to 0.75 for Qwen3-Embedding models (Y. Zhang et al. 2025) and 0.78 to 0.93 for similarly performing models. This suggests that random projection and neural training may capture comparable patterns, though SWeePtex-Emb lacks the fine-tuning capacity of larger neural models. The difference might arise from how neural networks explicitly learn biases, whereas random projections extract features without predefined assumptions. Bridging this gap could involve using random projections as initial embeddings before neural fine-tuning, reducing training costs while maintaining competitive performance.

Configuration analysis reveals several critical insights about the method's behavior. The comparable performance between train-only (560,000 documents) and full set (630,000 documents) demonstrates that test set inclusion in training data is not the primary driver of superior performance in larger models, as SWeePtex-Emb's full configuration gains no meaningful benefit from the additional test documents. However, the sample-size experiment shows an apparent growth trend with

some fluctuations across different training set sizes, without reaching a plateau at the limit tested, suggesting that larger corpora may improve results (Figure 2B). Combined with the observed performance plateau beyond 800 PCA dimensions (indicating that 1,200 dimensions provide sufficient feature coverage), these findings help delineate the method's fundamental capabilities. Additionally, the current tokenization method, simple whitespace splitting, may limit performance, and testing alternative approaches could yield further gains, although this is not explored in this study.

A limitation of comparative evaluation is the difficulty of quantifying modeling techniques for specialized domains, as exemplified by the Yoga study, which validated SWeePtex qualitatively but could not perform a quantitative comparison due to the lack of a supervised dataset. Nevertheless, qualitative validation remains valuable, and SWeePtex's applicability across domains invites future projects to demonstrate its versatility. Another experimental challenge is conducting fair comparisons with large models, such as top-performing LLMs on the MTEB, whose performance stems from massive, difficult-to-reproduce datasets. While retraining these models from the same architecture on a controlled dataset could enable a fairer test, this process is itself far from trivial.

Promising avenues include developing hybrid architectures that combine SWeePtex's random projection framework with lightweight neural components, potentially improving contextual task adaptation. By pursuing alternative development approaches and theoretical advancements in how random projections can capture linguistic structures, this research direction can offer a sustainable paradigm for representation learning, while promoting logical and arithmetical understanding of the process of embedding textual information.

5 Conclusion

This study demonstrates that SWeePtex-Emb achieves performance comparable to that of distilled small neural networks. The identified high correlation suggests that it captures foundational semantic patterns akin to those learned by deep learning approaches. These findings position it as a viable alternative for building specialized embedding models from scratch, particularly for domain-specific applications where data scarcity or computational constraints preclude the use of large pre-trained models. Crucially, the absence of a performance plateau with respect to the tested training data volume, combined with its ability to train on standard desktop hardware, indicates a clear direction for improvement through data scaling. While direct comparisons with Large Language Models (LLMs) involve inherent methodological complexities, the current results provide a foundation for further research into SWeePtex as a non-deep-learning paradigm for generating functional text embeddings.

References

De Pierri, C. R. et al. (Jan. 2020). "SWeeP: representing large biological sequences datasets in compact vectors". In: *Scientific Reports* 10.1, p. 91. ISSN: 2045-2322. DOI: 10.1038/s41598-019-55627-4. URL: https://doi/10.1038/s41598-019-55627-4 (visited on 02/01/2026).

DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. DOI: 10.48550/arXiv.2501.12948. URL: https://doi.org/10.48550/arXiv.2501.12948 (visited on 02/01/2026).

Hutson, Matthew (Sept. 2024). "Forget ChatGPT: why researchers now run small AIs on their laptops". In: *Nature* 633.8030, pp. 728–729. ISSN: 0028-0836. DOI: 10.1038/D41586-024-02998-Y. URL: https://doi.org/10.1038/D41586-024-02998-Y (visited on 02/01/2026).

Ieger-Raittz, R. et al. (May 2025). "What are we learning with Yoga? Mapping the scientific literature on Yoga using a vector-text-mining approach". In: *PLOS ONE* 20.5, e0322791. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0322791. URL: https://doi/10.1371/JOURNAL.PONE.0322791 (visited on 02/01/2026).

Kuratov, Yuri and Mikhail Arkhipov (2019). *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*. DOI: 10.48550/arXiv.1905.07213. URL: https://doi.org/10.48550/arXiv.1905.07213 (visited on 02/01/2026).

Lehmann, Jens et al. (2015). "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2, pp. 167–195. DOI: 10.3233/SW-140134. URL: https://doi.org/10.3233/SW-140134 (visited on 02/01/2026).

Li, Shen, Renfen Hu, and Lijun Wang (2025). *Efficiently Building a Domain-Specific Large Language Model from Scratch: A Case Study of a Classical Chinese Large Language Model*. DOI: 10.48550/arXiv.2505.11810. URL: https://doi.org/10.48550/arXiv.2505.11810 (visited on 02/01/2026).

Machado, D. J. S. et al. (Apr. 2021). "Biotext: Exploiting Biological-Text Format for Text Mining". In: *bioRxiv*. DOI: 10.1101/2021.04.08.439078. URL: https://doi/10.1101/2021.04.08.439078 (visited on 02/01/2026).

Muennighoff, N. et al. (Oct. 2022). "MTEB: Massive Text Embedding Benchmark". In: *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 2006–2029. DOI: 10.18653/v1/2023.eacl-main.148. URL: https://aclanthology.org/2023.eacl-main.148 (visited on 02/01/2026).

Perico, Camila P et al. (2022). "Genomic landscape of the SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin". In: *Frontiers in Microbiology* 13. DOI: 10.3389/fmicb.2022.1037455. URL: https://doi.org/10.3389/fmicb.2022.1037455 (visited on 02/01/2026).

Raittz, Roberto Tadeu et al. (2021). "Comparative genomics provides insights into the taxonomy of azoarcus and reveals separate origins of nif genes in the proposed azoarcus and aromatoleum genera". In: *Genes* 12, pp. 1–21. DOI: 10.3390/genes12010071. URL: https://doi.org/10.3390/genes12010071 (visited on 02/01/2026).

Schwartz, Roy et al. (Nov. 2020). "Green AI". In: *Commun. ACM* 63.12, pp. 54–63. ISSN: 0001-0782. DOI: 10.1145/3381831. URL: https://doi.org/10.1145/3381831 (visited on 02/01/2026).

Silva Filho, Antonio Camilo da et al. (2021). "Prediction and Analysis in silico of Genomic Islands in Aeromonas hydrophila". In: *Frontiers in Microbiology* 12. DOI: 10.3389/fmicb.2021.769380. URL: https://doi.org/10.3389/fmicb.2021.769380 (visited on 02/01/2026).

Yao, Xingcheng et al. (2022). *NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework*. DOI: 10.48550/arXiv.2111.04130. URL: https://doi.org/10.48550/arXiv.2111.04130 (visited on 02/01/2026).

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2016). *Character-level Convolutional Networks for Text Classification*. DOI: 10.48550/arXiv.1509.01626. URL: https://doi.org/10.48550/arXiv.1509.01626 (visited on 02/01/2026).

Zhang, Yanzhao et al. (2025). *Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models*. DOI: 10.48550/arXiv.2506.05176. URL: https://doi.org/10.48550/arXiv.2506.05176 (visited on 02/01/2026).

2.3 ARTIGO 3 – TXTREE: FERRAMENTA VISUAL PARA EXPLORAÇÃO DE LITERATURA

O Artigo 3 apresenta o TXTree, uma ferramenta em linha de comando que converte resultados de buscas do PubMed no formato MEDLINE em uma interface visual interativa e portátil denominada HTML-TM (*HyperText Markup Language for Text Mining*). Seu objetivo é facilitar a exploração e a mineração da literatura científica de forma acessível, reproduzível e operável localmente, demonstrando a funcionalidade do SWeePtex em forma de aplicativo.

O HTML-TM é previamente apresentado no estudo publicado sobre Yoga (Ieger-Raittz et al., 2025). Neste artigo, o foco está no resultado da exploração de literatura e nas indagações sobre as informações identificadas, constituindo uma demonstração do uso do SWeePtex em um problema real. Em continuidade, o Artigo 3 concentra-se na conceituação do HTML-TM em uma versão aprimorada e na apresentação da aplicação de geração, o TXTree, detalhando sua implementação e suas funcionalidades.

O processo central utiliza o método SWeePtex para gerar incorporações vetoriais de texto, que fundamentam a organização lógica dos documentos e dos termos, bem como o cálculo de similaridades na interface. A ferramenta gera dois módulos HTML principais: WORDS.html, para navegação por termos e suas relações, e TEXTS.html, para exploração de artigos científicos e de seus metadados.

Em comparação com ferramentas similares, como o SWIFT-Review e o TopicTracker, o TXTree apresenta características distintivas. Entre elas, destacam-se a geração de saída em HTML autossuficiente, a exportação dos vetores SWeePtex para análises computacionais subsequentes e a operação em modo *offline*, que dispensa qualquer registro de usuário e dependência de infraestrutura remota.

Um experimento com um corpus de literatura sobre doenças tropicais negligenciadas demonstra a aplicação prática da ferramenta. A navegação visual permite identificar associações semânticas entre termos e explorar literatura correlata. Adicionalmente, a funcionalidade de exportação de vetores viabiliza a análise programática externa, combinando técnicas de aprendizado de máquina não supervisionado com modelos de linguagem leves para a identificação automática de padrões e tendências de pesquisa.

Por meio de sua interface visual, o TXTree posiciona-se como uma solução acessível para usuários de domínios específicos não especialistas em programação. Também oferece, por meio da exportação de seus vetores, flexibilidade para pesquisadores aptos a programar, possibilitando a integração de *pipelines* personalizados na análise visual. A ferramenta está disponível publicamente para uso imediato, tanto

para revisão de literatura quanto como base para o desenvolvimento de aplicações futuras.

TXTree: A Visual Tool for PubMed Literature Exploration by Text Mining

Diogo de J. S. Machado[1], Flávia de F. Costa[1], Leonardo Vicenzi[1], Fábio de O. Pedrosa[1], and Roberto T. Raittz[1]

[1]Federal University of Paraná (UFPR)

Abstract

The current volume of scientific literature demands efficient tools for knowledge exploration. Current Artificial Intelligence (AI) solutions often suffer from limited portability, a lack of integrated features, or dependence on cloud services, which hinder both performance and data control. To address these gaps, this study presents TXTree (Text Tree), a command-line tool that transforms PubMed MEDLINE files into two complementary outputs: the HTML-TM (HyperText Markup Language for Text Mining), a portable, client-side platform for visually navigating and identifying associations between words and documents; and exported vector embeddings, that enable the development of complex, customized analytical pipelines. These two modes can be integrated to perform comprehensive research. The tool's effectiveness is demonstrated through an application to neglected tropical disease literature, where it successfully identified key research themes both visually and computationally. By combining advanced AI with a dual-purpose interface and operating entirely locally, TXTree offers a robust, self-contained framework that bridges a critical gap in literature-mining tools, ensuring full user control over data and methodology. The TXTree command-line tool is freely available on SourceForge (`https://sf.net/p/txtree`).

Keywords: Text mining. Literature exploration. Visual tool.

1 Introduction

The exponential growth of scientific literature presents a significant challenge, making it desirable to develop tools for efficient exploration and knowledge discovery. Researchers increasingly face the challenge of navigating vast, complex volumes of textual data to identify meaningful patterns, emerging trends, and non-obvious connections. This necessity drives the creation of platforms that process raw textual information and transform it into interactive, insightful, and intuitively navigable visual representations, thus augmenting human capacity for synthesis and analysis.

Although large language models (LLMs) offer broad applicability for general-purpose natural language processing (NLP), their deployment in rigorous academic contexts poses inherent challenges. The primary concern remains their tendency to generate superficially coherent and persuasive, yet fundamentally misleading or inaccurate results, a phenomenon often described as "hallucination" (Currie 2023). Furthermore, the academic research cycle places high value on the principles of data control, methodological reproducibility, and privacy, which become challenging to guarantee when relying on external, proprietary API-based models (Hutson 2024). These concerns are compounded by the substantial computational resource costs associated with the use of large-scale LLM, raising essential questions about the long-term sustainability of such approaches (Schwartz et al. 2020).

In response to these challenges, we present TXTree (text tree). Building on a practical and transparent methodology for knowledge exploration introduced in our previous study on Yoga (Ieger-Raittz et al. 2025), which is based on text vectorization and arithmetic operations, TXTree functions as a command-line tool engineered to automate the transformation of PubMed search results, delivered in standard MEDLINE format, directly into a fully functional visual navigation interface termed HTML-TM (HyperText Markup Language for Text Mining). The core computational methodology relies on generating text embeddings using the SWeeP random projection approach (De Pierri et al. 2020; Machado et al. 2021), offering an alternative to LLMs that is theoretically grounded in the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984).

A central feature of TXTree's design is its dual-interface utility, which supports both an intuitive visual navigation environment for domain experts and advanced programmatic access through the export of generated vectors for custom computational analysis. Crucially, this architecture enables substantial AI analyses to be performed locally on standard consumer hardware, eliminating dependence on cloud infrastructure and ensuring complete methodological transparency, reproducibility, and data control.

This approach is validated through a proof-of-concept analysis of the literature on neglected tropical diseases. TXTree is implemented as a Python-based command-line application and is also distributed as pre-compiled binaries for major operating systems. To contextualize its specific contributions, the following section presents a

focused review of the mining landscape in the existing literature and a comparative analysis with closely related tools.

1.1 Landscape of Literature Mining Tools

Within the ecosystem of visual, AI-driven tools for processing scientific literature in biomedicine, several functional categories can be identified, acknowledging that overlaps may occur. Using the PubMed database (Canese and Weis 2002) as a central resource, these categories include applications dedicated to the discovery of relationships between biomedical entities such as genes, diseases, and drugs (Dai et al. 2016; Jang et al. 2006; Macnee et al. 2021; Metzger et al. 2024; Theodosiou et al. 2008; Tsuruoka et al. 2011); annotation tools focused on the automated tagging and curation of precise biomedical concepts within text (Griffith et al. 2008; Neves 2020; Wei et al. 2019); integrative data platforms that combine text-mined information with multiomics data to provide a unified biological context (Hoffmann et al. 2014; H. Huang et al. 2018; Kafkas, Dunham, and McEntyre 2017; Kuo, Tian, and Tseng 2013); specialized databases that function as focused repositories for niches within biomedical science (Abulaish, Parwez, and Jahiruddin 2019; Aziz et al. 2024; Y. Huang, Wang, and Zan 2016; Jaylet et al. 2023; Li et al. 2022; Tudor et al. 2015; Turina, Fariselli, and Capriotti 2021; Yang et al. 2010); and knowledge exploration tools designed to support literature navigation and hypothesis generation through direct visual interaction (Chen et al. 2017; Gobeill et al. 2020; Howard et al. 2016; Ivanisenko et al. 2020; Nováček and Burns 2014; Rani, Shah, and Ramachandran 2015; Smalheiser, Fragnito, and Tirk 2021; Spitale, Germani, and Biller-Andorno 2024).

TXTree falls primarily into the category of knowledge exploration tools, which are explicitly designed to support the exploration of the literature through direct visual interaction with corpora (Chen et al. 2017; Gobeill et al. 2020; Howard et al. 2016; Ivanisenko et al. 2020; Nováček and Burns 2014; Rani, Shah, and Ramachandran 2015; Smalheiser, Fragnito, and Tirk 2021; Spitale, Germani, and Biller-Andorno 2024). From a focused search for comparable systems, we identified SWIFT-Review (Howard et al. 2016) and TopicTracker (Spitale, Germani, and Biller-Andorno 2024) as the most directly related and currently available tools for comparison. Another tool, Anne O'Tate (Smalheiser, Fragnito, and Tirk 2021), was identified in the literature but could not be tested due to its apparent unavailability.

A feature-based comparison highlights TXTree's distinct position in this niche (Table 1). While SWIFT-Review provides valuable automated topic categorization through a graphical user interface, and TopicTracker enables frequency-based analysis and graph generation within computational notebooks, TXTree establishes a different paradigm. Its core differentiators are the generation of a portable, self-contained HTML interface that requires no installation, coupled with full machine learning readiness through the export of SWeePtex word and document vectors.

This combination prioritizes a workflow in which intuitive visual exploration seamlessly transitions into programmatic, AI-enhanced analysis, all within a transparent, offline-capable ecosystem.

Table 1: Theoretical comparison of TXTree with similar tools (SWIFT-Review and TopicTracker). The table contrasts TXTree's capabilities with two other scientific literature mining tools, highlighting their distinct approaches to text processing, data visualization, and machine learning integration. Symbols (✓ = supported, × not supported, ∼ = partially/limited support) objectively indicate the presence or absence of specific features in each tool. The comparison covers vector export, topic modeling, output formats, and system requirements, serving as a technical reference for positioning TXTree within the ecosystem of text analysis tools.

Feature	**SWIFT-Review**	**TopicTracker**	**TXTree**
Release year	2016	2024	2025
Vector Export	× No vector output	∼ Frequency-based counts	✓ Word and document vectors (SWeePtex)
ML Readiness	× Not applicable	∼ Frequency-based models	✓ Full (SWeePtex vectors + PCA space)
Topic Modeling	✓ Automatic categorization	∼ Manual cooccurrence analysis	∼ Externally
Interface	✓ Graphical UI	× Jupyter Notebooks only	✓ Interactive HTML
Output Format	× Proprietary binary	✓ CSV / Notebooks	✓ Portable HTML
Related Concepts	✓ Topic groups	∼ Manual co-occurrence	✓ Semantic links (words + documents)
Intelligent Ordering	✓ Score-based within categories	× Manual sorting required	✓ Cosine similarity + dendrogram
Own Navigation Interface	✓ Dedicated GUI	× Relies on external tools	✓ HTML-TM with search/sort
Portable Output Format	× Proprietary format	✓ Open formats (CSV / IPYNB)	✓ Self-contained HTML
Offline Use	✓ With PubMed XML	✓ With MEDLINE	✓ With MEDLINE
Registration-free	× No	✓ Yes	✓ Yes
Best For	Quick topic categorization	Graph generation and bibliometric plots	Visual literature mining with advanced usage for programmers

2 Method

TXTree developing and validating consists of two key components: 1) a description of its technical implementation; and 2) a usage example demonstrating its functionality.

The technical implementation of TXTree is a specialized command-line tool that processes MEDLINE/PubMed data files to produce HTML-TM output. The system consists of two core modules: a lexical overview (WORDS.html) and a document-level explorer (TEXTS.html). Each module provides access to retrieved publications, featuring intelligent content ordering and an integrated search system. The implementation uses Python 3 (Van Rossum and Foundation 2026) and is compiled with Nuitka (Hayen and Contributors 2025). The design incorporates NumPy (Harris et al. 2020) and Pandas (McKinney 2010) for data handling, Scikit-learn (Pedregosa et al. 2011) for machine learning workflows, ETE3 (Huerta-Cepas, Serra,

and Bork 2016) and Matplotlib (Hunter 2007) for plotting, and Biopython (Cock et al. 2009) for FASTA file processing as required by the Biotext framework.

The usage example demonstrates the practical utility of TXTree by analyzing the literature on "neglected tropical diseases" with a specific focus on trypanosomatids. The example uses the results of a PubMed search from 26 March 2025, using the query:

```
neglected[title] AND (diseases[title/abstract] OR disease[title/
abstract]) AND english[language] AND hasabstract NOT "Published Erratum
"[Publication Type]
```

TXTree is designed to run with the results of this search using the following command:

```
TXTree --html_tm_title "Neglected Diseases" --save_emb --output_dir
neglected_diseases_html_tm neglected_diseases.medline
```

The `--save_emb` flag enables the export of embedding vectors, PCA projection vectors, and additional data to an HDF5 file for subsequent programmatic analysis by external AI systems.

The analysis is demonstrated through human reading and integration with an LLM-based chatbot via the DeepSeek (DeepSeek-AI et al. 2025) web interface. The interaction is conducted through three prompts. The first is a simple request to submit a word to generate a contextual paragraph. The second instructs the model to place the term "molecules" within its expected context. The third is a request to incorporate the concept of drug repositioning.

The programmatic interface demonstration processes the document embeddings and PCA vectors from the exported HDF5 file, filtering documents related to trypanosomatids using a regular expression ("trypa.*"). In this pipeline, the elbow method determines the optimal number of clusters for subsequent k-means clustering based on PCA vectors. A power-law-based sampling strategy selects representative documents from each resulting cluster. Finally, a lightweight LLM – specifically the Qwen3 1.7B model (Zhang et al. 2025) deployed via Ollama (Ollama 2026) – processes these documents using a predefined prompt template that instructs the model to analyze titles and generate concise thematic summaries for each cluster.

3 Result

TXTree is implemented as a command-line tool designed for comprehensive literature exploration. Its AI-based processing pipeline takes a MEDLINE file from a PubMed search as input and generates an interactive HTML-TM file, which serves as a visual interface for analysis (Figure 1). The software is freely available on

SourceForge[1], which provides precompiled versions for Windows and Linux. The source code can be found on PyPI[2]. A file with usage instructions and examples of commands for different situations is also available on SourceForge.

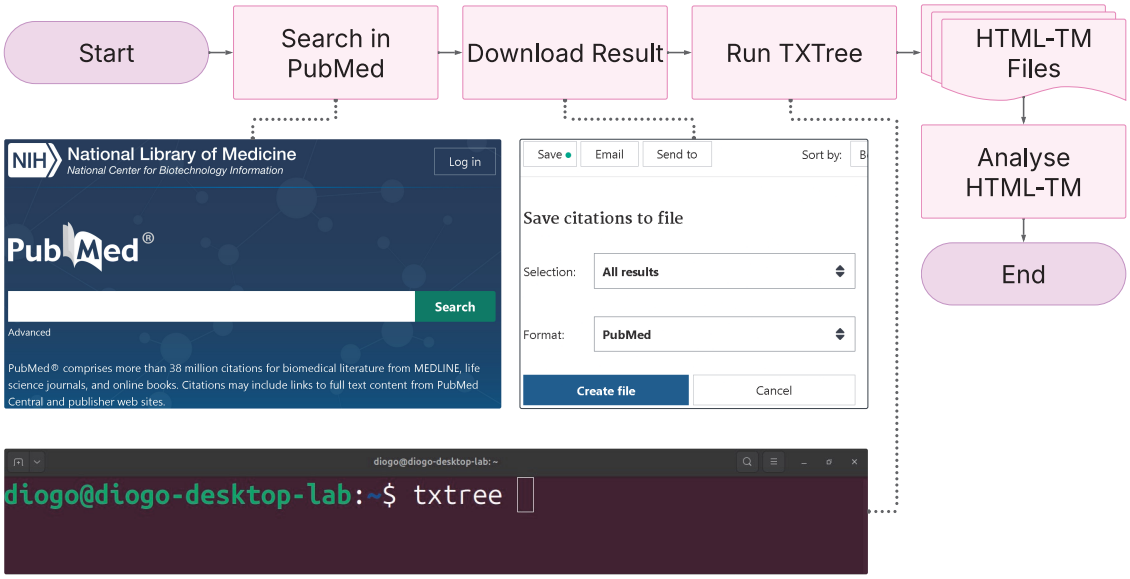


Figure 1: Pipeline for HTML-TM file generation and analysis. The process starts with a PubMed literature search, followed by downloading the results and processing them with TXTree to generate the HTML-TM file, which culminates in analyzing the output.

The WORDS.html and TEXTS.html files are created with a table with the corresponding data as the main element; in the first, each row represents a word in the word list, while in the second, each row represents a document (paper) in the corpus. Furthermore, the Related Documents pages, accessible from the previous files, provide information about documents associated with a specific word or paper based on cosine similarity. Table 2 outlines the structure of the file tables. HTML-TM pages are improved with a search utility that supports regular expressions and column-specific queries. The dendrogram-based organization of documents and terms within the interface facilitates logical exploration of a given literature corpus. The interactive navigation process enables users to quickly filter and explore key terms and their surrounding context.

The usage example demonstrates the application of TXTree to the theme of "neglected diseases", starting with a PubMed search that yields a MEDLINE file containing 2,471 entries. Processing through TXTree generates WORDS.html with 7,252 entries and TEXTS.html with 2,471 entries. Figure 2A displays the WORDS.html page with an active filter set for the target word "trypanosomatids". In the corresponding row, the related words column includes trypanosomatids, molecules, try-

[1] `https://sf.net/p/txtree`
[2] `https://pypi.org/p/txtree`

Table 2: HTML-TM columns. Structure of the files W (WORDS.html), T (TEXTS.html), and R (Related Documents). An X marks the columns in each file, while blank cells indicate the absence of the column. The table compares the columns across the three files, including their descriptions.

Column	W	T	R	Description
ID	X	X	X	Identifies each word or paper in dendrogrammatic order.
Word	X			Displays the words.
Occ.	X			Indicates the frequency of each word in the corpus.
Related Words	X			Lists the ten closest words (defined by cosine similarity).
Related Doc.	X	X		Provides a link to related documents (defined by cosine similarity).
Word Tree	X			Offers a link to a dendrogram displaying related words.
Year Plot	X			Links to a temporal correlation plot showing the word's usage over time.
Title		X		States the title of the paper.
Year		X	X	Indicates the publication year.
PMID		X	X	PubMed Identifier hyperlinked to the PubMed page.
Rank			X	Defines the similarity in descending order (defined by cosine similarity).
Similarity			X	Provides the cosine similarity with the query row (word or document).
Title + Abs			X	Lists the documents (titles and abstracts concatenated) with the paper title bolded.

panosomatid, biologically, organisms, protozoan, chemical, chemotherapeutic, trypanosoma, and parasites. The closest term to "trypanosomatids" is "molecules", which is used as a filter on the related documents page (Figure 2B). The neighboring terms, also in Figure 2A, include throughput, silico, repositioning, eukaryotic, trypanosomatid, trypanosomatids, isatin, display, trypanosomal, isoniazid, and antituberculosis.

The written text to demonstrate the logical order of neighboring words is available in the Supplementary Box S1. The narrative starts with high-throughput screening (HTS) as a foundational method for drug discovery, then transitions into the roles of in silico approaches and drug repositioning strategies. The text subsequently narrows its focus to eukaryotic targets within trypanosomatids, discusses specific compounds such as isatin and isoniazid, and concludes by highlighting the broad potential of hybrid compounds and repurposed drugs.

To illustrate TEXT.html, Figure 3A shows a search for the PMID of the article most closely associated with the target word and its neighboring terms. The subsequent Figure 3B presents the top six results of the related documents page for that same article.

The interaction with DeepSeek is illustrated in Supplementary Box S2. The final response successfully integrates the specific context of therapeutic molecules. It also introduces the strategic approach of drug repositioning, aligning with the research focus on developing chemotherapeutics for trypanosomatid infections.

In the programmatic usage demonstration, an unsupervised machine learning technique (clustering) is applied to the exported PCA vectors to identify latent patterns and group similar documents. Subsequently, these structured clusters are processed by a lightweight, locally run LLM (Qwen3 1.7B) to generate thematic summaries. The raw output of this automated theme extraction process is available in the Supplementary Table S1.

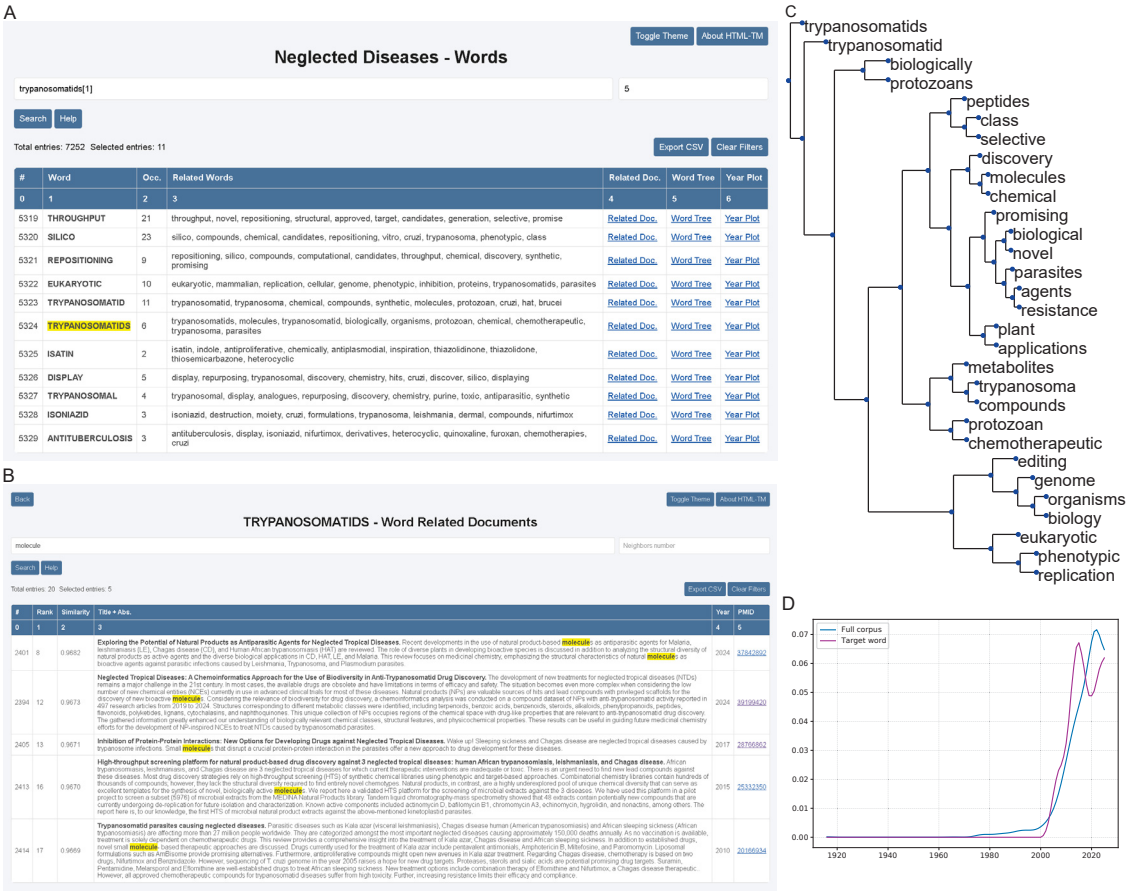
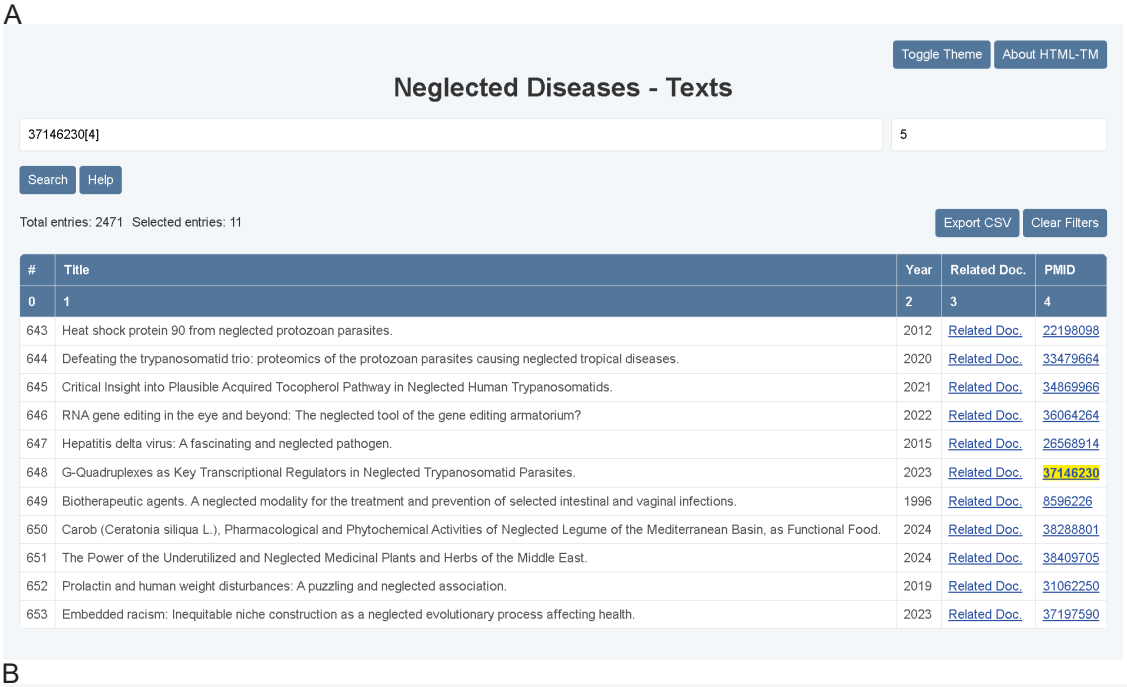


Figure 2: WORDS.html navigation prints. (A) Example of HTML-TM for the topic "neglected diseases" with a focus on "trypanosomatids". The target term "trypanosomatids[1]" is searched in the WORDS.html text box, where "[1]" indicates the search column (Word column). The "Neighbors number" is set to 5, retrieving ten items (five before and five after the matching term) in addition to the match itself. The results are displayed in the figure. (B) Filtered results for the term "molecule" in trypanosomatids-related documents. (C) Word Tree for the target word, displaying dendrogrammatic relationships of related terms. (D) Year Plot for the target word, showing temporal trends in the word's usage across the corpus.

The automated thematic analysis (Supplementary Table S1) identifies sixteen research clusters (0-15). Key strategies involve exploring natural therapeutics such as plant-derived triterpenes (0) and essential oils (2), developing synthetic compounds (5), and creating hybrid molecules (3). It targets specific biological pathways, including kinases (10), mitochondrial functions (12), and aminoacyl-tRNA synthetases (14). Methodologically, the corpus includes structural-activity relationship studies (13), computational modeling (6), drug repurposing efforts (10), and nanopharmaceutical systems (15). The analysis also covers molecular mechanisms of disease progression (1), translational challenges in trypanosome diseases (9), and ecological transmission involving vectors (4, 11).

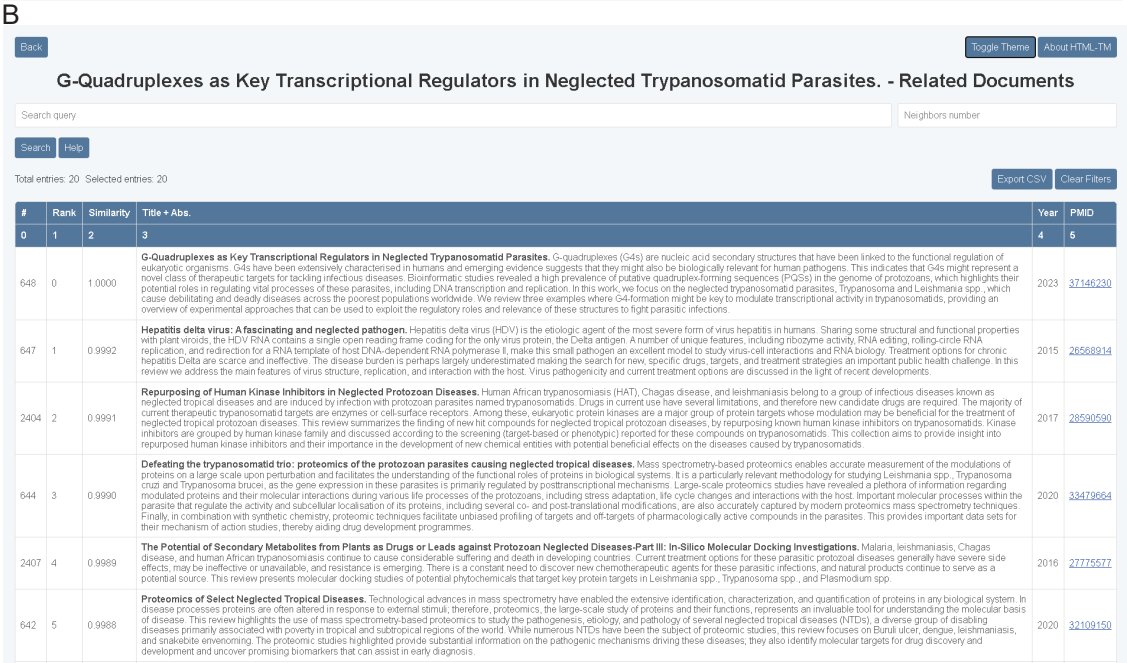
The HTML-TM generated from the usage example, the script used for external

A



#	Title	Year	Related Doc.	PMID
0	1	2	3	4
643	Heat shock protein 90 from neglected protozoan parasites.	2012	Related Doc.	22198098
644	Defeating the trypanosomatid trio: proteomics of the protozoan parasites causing neglected tropical diseases.	2020	Related Doc.	33479664
645	Critical Insight into Plausible Acquired Tocopherol Pathway in Neglected Human Trypanosomatids.	2021	Related Doc.	34869966
646	RNA gene editing in the eye and beyond: The neglected tool of the gene editing armatorium?	2022	Related Doc.	36064264
647	Hepatitis delta virus: A fascinating and neglected pathogen.	2015	Related Doc.	26568914
648	G-Quadruplexes as Key Transcriptional Regulators in Neglected Trypanosomatid Parasites.	2023	Related Doc.	37146230
649	Biotherapeutic agents. A neglected modality for the treatment and prevention of selected intestinal and vaginal infections.	1996	Related Doc.	8596226
650	Carob (Ceratonia siliqua L.), Pharmacological and Phytochemical Activities of Neglected Legume of the Mediterranean Basin, as Functional Food.	2024	Related Doc.	38288801
651	The Power of the Underutilized and Neglected Medicinal Plants and Herbs of the Middle East.	2024	Related Doc.	38409705
652	Prolactin and human weight disturbances: A puzzling and neglected association.	2019	Related Doc.	31062250
653	Embedded racism: Inequitable niche construction as a neglected evolutionary process affecting health.	2023	Related Doc.	37197590

B



#	Rank	Similarity	Title + Abs.	Year	PMID
0	1	2	3	4	5
648	0	1.0000	**G-Quadruplexes as Key Transcriptional Regulators in Neglected Trypanosomatid Parasites.** G-quadruplexes (G4s) are nucleic acid secondary structures that have been linked to the functional regulation of eukaryotic organisms. G4s have been extensively characterised in humans and emerging evidence suggests that they might also be biologically relevant for human pathogens. This indicates that G4s might represent a novel class of therapeutic targets for tackling infectious diseases. Bioinformatic studies revealed a high prevalence of putative quadruplex-forming sequences (PQSs) in the genome of protozoans, which highlights their potential roles in regulating vital processes of these parasites, including DNA transcription and replication. In this work, we focus on the neglected trypanosomatid parasites, Trypanosoma and Leishmania spp., which cause debilitating and deadly diseases across the poorest populations worldwide. We review three examples where G4-formation might be key to modulate transcriptional activity in trypanosomatids, providing an overview of experimental approaches that can be used to exploit the regulatory roles and relevance of these structures to fight parasitic infections.	2023	37146230
647	1	0.9992	**Hepatitis delta virus: A fascinating and neglected pathogen.** Hepatitis delta virus (HDV) is the etiologic agent of the most severe form of virus hepatitis in humans. Sharing some structural and functional properties with plant viroids, the HDV RNA contains a single open reading frame coding for the only virus protein, the Delta antigen. A number of unique features, including ribozyme activity, RNA editing, rolling-circle RNA replication, and redirection for a RNA template of host DNA-dependent RNA polymerase II, make this small pathogen an excellent model to study virus-cell interactions and RNA biology. Treatment options for chronic hepatitis Delta are scarce and ineffective. The disease burden is perhaps largely underestimated making the search for new, specific drugs, targets, and treatment strategies an important public health challenge. In this review we address the main features of virus structure, replication, and interaction with the host. Virus pathogenicity and current treatment options are discussed in the light of recent developments.	2015	26568914
2404	2	0.9991	**Repurposing of Human Kinase Inhibitors in Neglected Protozoan Diseases.** Human African trypanosomiasis (HAT), Chagas disease, and leishmaniasis belong to a group of infectious diseases known as neglected tropical diseases and are induced by infection with protozoan parasites named trypanosomatids. Drugs in current use have several limitations, and therefore new candidate drugs are required. The majority of current therapeutic trypanosomatid targets are enzymes or cell-surface receptors. Among these, eukaryotic protein kinases are a major group of protein targets whose modulation may be beneficial for the treatment of neglected tropical protozoan diseases. This review summarizes the finding of new hit compounds for neglected tropical protozoan diseases, by repurposing known human kinase inhibitors on trypanosomatids. Kinase inhibitors are grouped by human kinase family and discussed according to the screening (target-based or phenotypic) reported for these compounds on trypanosomatids. This collection aims to provide insight into repurposed human kinase inhibitors and their importance in the development of new chemical entities with potential beneficial effects on the diseases caused by trypanosomatids.	2017	28590590
644	3	0.9990	**Defeating the trypanosomatid trio: proteomics of the protozoan parasites causing neglected tropical diseases.** Mass spectrometry-based proteomics enables accurate measurement of the modulations of proteins on a large scale upon perturbation and facilitates the understanding of the functional roles of proteins in biological systems. It is a particularly relevant methodology for studying Leishmania spp., Trypanosoma cruzi and Trypanosoma brucei, as the gene expression in these parasites is primarily regulated by posttranscriptional mechanisms. Large-scale proteomics studies have revealed a plethora of information regarding modulated proteins and their molecular interactions during various life processes of the protozoans, including stress adaptation, life cycle changes and interactions with the host. Important molecular processes within the parasite that regulate the activity and subcellular localisation of its proteins, including several co- and post-translational modifications, are also accurately captured by modern proteomics mass spectrometry techniques. Finally, in combination with synthetic chemistry, proteomic techniques facilitate unbiased profiling of targets and off-targets of pharmacologically active compounds in the parasites. This provides important data sets for their mechanism of action studies, thereby aiding drug development programmes.	2020	33479664
2407	4	0.9989	**The Potential of Secondary Metabolites from Plants as Drugs or Leads against Protozoan Neglected Diseases-Part III: In-Silico Molecular Docking Investigations.** Malaria, leishmaniasis, Chagas disease, and human African trypanosomiasis continue to cause considerable suffering and death in developing countries. Current treatment options for these parasitic protozoal diseases generally have severe side effects, may be ineffective or unavailable, and resistance is emerging. There is a constant need to discover new chemotherapeutic agents for these parasitic infections, and natural products continue to serve as a potential source. This review presents molecular docking studies of potential phytochemicals that target key protein targets in Leishmania spp., Trypanosoma spp., and Plasmodium spp.	2016	27775577
642	5	0.9988	**Proteomics of Select Neglected Tropical Diseases.** Technological advances in mass spectrometry have enabled the extensive identification, characterization, and quantification of proteins in any biological system. In disease processes proteins are often altered in response to external stimuli, therefore, proteomics, the large-scale study of proteins and their functions, represents an invaluable tool for understanding the molecular basis of disease. This review highlights the use of mass spectrometry-based proteomics to study the pathogenesis, etiology, and pathology of several neglected tropical diseases (NTDs), a diverse group of disabling diseases primarily associated with poverty in tropical and subtropical regions of the world. While numerous NTDs have been the subject of proteomic studies, this review focuses on Buruli ulcer, dengue, leishmaniasis, and snakebite envenoming. The proteomic studies highlighted provide substantial information on the pathogenic mechanisms driving these diseases; they also identify molecular targets for drug discovery and development and uncover promising biomarkers that can assist in early diagnosis.	2020	32109150

Figure 3: TEXTS.html navigation prints. (A) PMID search results for the article most associated with "trypanosomatids" and its neighboring terms. (B) The top six related documents retrieved for the PMID highlighted in (A).

programmatic analysis, and their outputs are hosted on Zenodo[3].

[3] `https://doi.org/10.5281/zenodo.18370865`

4 Discussion

TXTree operationalizes the concept of portable knowledge exploration by transforming static PubMed datasets into an interactive HTML-TM interface. This approach directly addresses the challenges of accessibility and methodological transparency outlined in the Introduction. The HTML-TM output, viewable in any standard web browser, democratizes advanced literature analysis by eliminating the need for software installation, server infrastructure, or specialized computational skills. Its client-side operation ensures data privacy and enables offline use, which are critical for sensitive research projects or environments with restricted internet access.

The utility of this approach is demonstrated through an example analyzing the neglected tropical disease literature, which successfully uncovers coherent research themes such as drug repositioning and plant-based therapeutics. Our previous publication on Yoga validates the methodological foundation, which established the core HTML-TM framework and its application for literature exploration (Ieger-Raittz et al. 2025). Together, these studies provide qualitative validation of the random projection approach implemented via SWeeP as an effective text embedding technique for exploratory analysis. Building on this validated foundation, TXTree automates the pipeline and extends its capabilities with programmatic access to machine-learning-ready data.

This unified architecture enables the two-stage AI pipeline demonstrated in the results: exported vectors can be processed with traditional machine learning techniques, such as clustering, and the resulting structured data can feed into a locally run LLM for summarization. By enabling this pipeline on consumer hardware, TXTree provides a practical alternative to cloud-dependent AI tools, mitigating concerns about cost, data control, and reproducibility associated with external API-based models (Hutson 2024; Schwartz et al. 2020). The usage example demonstrates how traditional AI approaches can be integrated with current generative LLM techniques, even with lightweight models running locally.

Future development of TXTree can focus on optimizing its resource efficiency for massive corpora, a known trade-off of its current comprehensive client-side design. Other potential enhancements include exploring modular extensions to its HTML-TM interface, such as integrated basic statistical charts or user annotation layers. As an openly available tool, TXTree serves as both an immediate solution for literature exploration and a foundational module for building customized, transparent, and portable text-mining workflows.

5 Conclusion

TXTree provides a versatile, self-contained solution for literature reviews by transforming PubMed datasets into an interactive HTML-TM platform. Its dual-

interface design successfully bridges two research modalities: an intuitive, visual explorer for domain experts and a programmatic gateway via exported vectors for advanced computational analysis. By prioritizing portability, transparency, and user control, TXTree provides a practical foundation for navigating the expanding scientific corpus, enabling both immediate visual discovery and deeper, AI-ready investigation.

References

Abulaish, Muhammad, Md Aslam Parwez, and Jahiruddin (Dec. 2019). *DiseaSE: A biomedical text analytics system for disease symptom extraction and characterization*. en. DOI: 10.1016/j.jbi.2019.103324. URL: https://doi.org/10.1016/j.jbi.2019.103324 (visited on 02/01/2026).

Aziz, Muzzamil et al. (Nov. 2024). *KnowVID-19: A knowledge-based system to extract targeted COVID-19 information from online medical repositories*. en. DOI: 10.3390/biom14111411. URL: https://doi.org/10.3390/biom14111411 (visited on 02/01/2026).

Canese, Kathi and Sarah Weis (2002). *PubMed: The Bibliographic Database*. URL: https://www.ncbi.nlm.nih.gov/books/NBK153385/pdf/Bookshelf_NBK153385.pdf (visited on 02/01/2026).

Chen, Qian et al. (Sept. 2017). *Revealing topics and their evolution in biomedical literature using Bio-DTM: a case study of ginseng*. en. DOI: 10.1186/s13020-017-0148-7. URL: https://doi.org/10.1186/s13020-017-0148-7 (visited on 02/01/2026).

Cock, Peter J A et al. (2009). *Biopython: freely available Python tools for computational molecular biology and bioinformatics*. DOI: 10.1093/bioinformatics/btp163. URL: https://doi.org/10.1093/bioinformatics/btp163 (visited on 02/01/2026).

Currie, Geoffrey M (2023). "Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?" In: *Seminars in Nuclear Medicine* 53.5, pp. 719–730. DOI: 10.1053/j.semnuclmed.2023.04.008. URL: https://doi.org/10.1053/j.semnuclmed.2023.04.008 (visited on 02/01/2026).

Dai, Hong-Jie et al. (May 2016). *MET network in PubMed: a text-mined network visualization and curation system*. en. DOI: 10.1093/database/baw090. URL: https://doi.org/10.1093/database/baw090 (visited on 02/01/2026).

De Pierri, Camilla Reginatto et al. (Jan. 2020). "SWeeP: representing large biological sequences datasets in compact vectors". In: *Scientific Reports* 10.1, p. 91. ISSN: 2045-2322. DOI: 10.1038/s41598-019-55627-4. URL: https://doi.org/10.1038/s41598-019-55627-4 (visited on 02/01/2026).

DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. DOI: 10.48550/arXiv.2501.12948. URL: https://doi.org/10.48550/arXiv.2501.12948 (visited on 02/01/2026).

Gobeill, Julien et al. (July 2020). *SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts*. en. DOI: 10.1093/nar/gkaa328. URL: https://doi.org/10.1093/nar/gkaa328 (visited on 02/01/2026).

Griffith, Obi L et al. (Jan. 2008). *ORegAnno: an open-access community-driven resource for regulatory annotation*. en. DOI: 10.1093/nar/gkm967. URL: https://doi.org/10.1093/nar/gkm967 (visited on 02/01/2026).

Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature 2020 585:7825* 585.7825, pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (visited on 02/01/2026).

Hayen, Kay and Nuitka Contributors (2025). *Nuitka the Python Compiler*. URL: https://nuitka.net (visited on 02/01/2026).

Hoffmann, Michael F et al. (Jan. 2014). *The Transformer database: biotransformation of xenobiotics*. en. DOI: 10.1093/nar/gkt1246. URL: https://doi.org/10.1093/nar/gkt1246 (visited on 02/01/2026).

Howard, Brian E et al. (May 2016). *SWIFT-Review: a text-mining workbench for systematic review*. en. DOI: 10.1186/s13643-016-0263-z. URL: https://doi.org/10.1186/s13643-016-0263-z (visited on 02/01/2026).

Huang, Hongzhan et al. (Jan. 2018). *iPTMnet: an integrated resource for protein post-translational modification network discovery*. DOI: 10.1093/nar/gkx1104. URL: https://doi.org/10.1093/nar/gkx1104 (visited on 02/01/2026).

Huang, Yan, Li Wang, and Lin-Sen Zan (Dec. 2016). *ARN: Analysis and visualization system for adipogenic regulation network information*. en. DOI: 10.1038/srep39347. URL: https://doi.org/10.1038/srep39347 (visited on 02/01/2026).

Huerta-Cepas, Jaime, François Serra, and Peer Bork (2016). "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". In: *Molecular Biology and Evolution* 33, pp. 1635–1638. DOI: 10.1093/molbev/msw046. URL: https://doi.org/10.1093/molbev/msw046 (visited on 02/01/2026).

Hunter, John D. (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55. URL: https://doi.org/10.1109/MCSE.2007.55 (visited on 02/01/2026).

Hutson, Matthew (Sept. 2024). "Forget ChatGPT: why researchers now run small AIs on their laptops". In: *Nature* 633.8030, pp. 728–729. ISSN: 0028-0836. DOI: 10.1038/D41586-024-02998-Y. URL: https://doi.org/10.1038/D41586-024-02998-Y (visited on 02/01/2026).

Ieger-Raittz, Rosangela et al. (May 2025). "What are we learning with Yoga? Mapping the scientific literature on Yoga using a vector-text-mining approach". In: *PLOS ONE* 20.5, e0322791. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0322791. URL: https://doi.org/10.1371/JOURNAL.PONE.0322791 (visited on 02/01/2026).

Ivanisenko, Timofey V et al. (Sept. 2020). *ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature*. en. DOI: 10.1186/s12859-020-03557-8. URL: https://doi.org/10.1186/s12859-020-03557-8 (visited on 02/01/2026).

Jang, Hyunchul et al. (Aug. 2006). *BioProber: Software system for biomedical relation discovery from PubMed*. DOI: 10.1109/IEMBS.2006.259838. URL: https://doi.org/10.1109/IEMBS.2006.259838 (visited on 02/01/2026).

Jaylet, Thomas et al. (July 2023). *AOP-helpFinder 2.0: Integration of an event-event searches module*. en. DOI: 10.1016/j.envint.2023.108017. URL: https://doi.org/10.1016/j.envint.2023.108017 (visited on 02/01/2026).

Johnson, William B. and Joram Lindenstrauss (1984). "Extensions of Lipschitz mappings into a Hilbert space". In: *Contemporary Mathematics*. American Mathematical Society, pp. 189–206. DOI: 10.1090/conm/026/737400. URL: https://doi.org/10.1090/conm/026/737400 (visited on 02/01/2026).

Kafkas, Şenay, Ian Dunham, and Johanna McEntyre (Dec. 2017). *Literature evidence in open targets - a target validation platform*. DOI: 10.1186/s13326-017-0131-3. URL: https://doi.org/10.1186/s13326-017-0131-3 (visited on 02/01/2026).

Kuo, Tien-Chueh, Tze-Feng Tian, and Yufeng Jane Tseng (July 2013). *3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data*. en. DOI: 10.1186/1752-0509-7-64. URL: https://doi.org/10.1186/1752-0509-7-64 (visited on 02/01/2026).

Li, Jin et al. (Nov. 2022). *PlagueKD: a knowledge graph-based plague knowledge database*. en. DOI: 10.1093/database/baac100. URL: https://doi.org/10.1093/database/baac100 (visited on 02/01/2026).

Machado, Diogo de Jesus Soares et al. (Apr. 2021). "Biotext: Exploiting Biological-Text Format for Text Mining". In: *bioRxiv*, p. 2021.04.08.439078. DOI: 10.1101/2021.04.08.439078. URL: https://doi.org/10.1101/2021.04.08.439078 (visited on 02/01/2026).

Macnee, Marie et al. (Nov. 2021). *SimText: a text mining framework for interactive analysis and visualization of similarities among biomedical entities*. en. DOI: 10.1093/bioinformatics/btab365. URL: https://doi.org/10.1093/bioinformatics/btab365 (visited on 02/01/2026).

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

Metzger, Vincent T et al. (June 2024). *TIN-X version 3: update with expanded dataset and modernized architecture for enhanced illumination of understudied targets*. en. DOI: 10.7717/peerj.17470. URL: https://doi.org/10.7717/peerj.17470 (visited on 02/01/2026).

Neves, Mariana (June 2020). *Integration of the PubAnnotation ecosystem in the development of a web-based search tool for alternative methods*. en. DOI: 10.5808/GI.2020.18.2.e18. URL: https://doi.org/10.5808/GI.2020.18.2.e18 (visited on 02/01/2026).

Nováček, Vít and Gully A.P.C. Burns (July 2014). "SKIMMR: facilitating knowledge discovery in life sciences by machine-aided skim reading". In: *PeerJ* 2, e483. ISSN: 2167-8359. DOI: 10.7717/peerj.483. URL: http://doi.org/10.7717/peerj.483 (visited on 02/01/2026).

Ollama (2026). *Ollama's documentation*. URL: https://docs.ollama.com (visited on 02/01/2026).

Pedregosa, F et al. (2011). *Scikit-learn: Machine Learning in Python*. URL: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (visited on 02/01/2026).

Rani, Jyoti, A B Rauf Shah, and Srinivasan Ramachandran (Oct. 2015). *pubmed.mineR: an R package with text-mining algorithms to analyse PubMed abstracts*. en. DOI: 10.1007/s12038-015-9552-2. URL: https://doi.org/10.1007/s12038-015-9552-2 (visited on 02/01/2026).

Schwartz, Roy et al. (Nov. 2020). "Green AI". In: *Commun. ACM* 63.12, pp. 54–63. ISSN: 0001-0782. DOI: 10.1145/3381831. URL: https://doi.org/10.1145/3381831 (visited on 02/01/2026).

Smalheiser, Neil R, Dean P Fragnito, and Eric E Tirk (Mar. 2021). *Anne O'Tate: Value-added PubMed search engine for analysis and text mining*. en. DOI: 10.1371/journal.pone.0248335. URL: https://doi.org/10.1371/journal.pone.0248335 (visited on 02/01/2026).

Spitale, Giovanni, Federico Germani, and Nikola Biller-Andorno (Sept. 2024). *TopicTracker - An advanced software pipeline for text mining on PubMed data: Bridging the gap between off-the-shelf tools and code based approaches*.

en. DOI: 10.1016/j.heliyon.2024.e36351. URL: https://doi.org/10.1016/j.heliyon.2024.e36351 (visited on 02/01/2026).

Theodosiou, T et al. (Sept. 2008). *PuReD-MCL: a graph-based PubMed document clustering methodology*. en. DOI: 10.1093/bioinformatics/btn318. URL: https://doi.org/10.1093/bioinformatics/btn318 (visited on 02/01/2026).

Tsuruoka, Yoshimasa et al. (July 2011). *Discovering and visualizing indirect associations between biomedical concepts*. en. DOI: 10.1093/bioinformatics/btr214. URL: https://doi.org/10.1093/bioinformatics/btr214 (visited on 02/01/2026).

Tudor, Catalina O et al. (Mar. 2015). *Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system*. en. DOI: 10.1093/database/bav020. URL: https://doi.org/10.1093/database/bav020 (visited on 02/01/2026).

Turina, Paola, Piero Fariselli, and Emidio Capriotti (Mar. 2021). *ThermoScan: Semi-automatic identification of protein stability data from PubMed*. en. DOI: 10.3389/fmolb.2021.620475. URL: https://doi.org/10.3389/fmolb.2021.620475 (visited on 02/01/2026).

Van Rossum, Guido and Python Software Foundation (2026). *Python Language Reference*. URL: https://docs.python.org/3/reference (visited on 02/01/2026).

Wei, Chih-Hsuan et al. (July 2019). *PubTator central: automated concept annotation for biomedical full text articles*. en. DOI: 10.1093/nar/gkz389. URL: https://doi.org/10.1093/nar/gkz389 (visited on 02/01/2026).

Yang, Lun et al. (Aug. 2010). *ReCGiP, a database of reproduction candidate genes in pigs based on bibliomics*. en. DOI: 10.1186/1477-7827-8-96. URL: https://doi.org/10.1186/1477-7827-8-96 (visited on 02/01/2026).

Zhang, Y. et al. (2025). *Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models*. DOI: 10.48550/arXiv.2506.05176. URL: https://doi.org/10.48550/arXiv.2506.05176 (visited on 02/01/2026).

Supplementary Material

Box S1: Human-written neighboring words text. This box presents an example of reading based on a sequence of words in the HTML-TM (Hypertext Markup Language for Text Mining) listing. The words appear in sequence in the HTML-TM built with PubMed data related to "neglected diseases''. Explanations about HTML-TM can be found in the main document. The words are: throughput, silico, repositioning, eukaryotic, trypanosomatid, trypanosomatids, isatin, display, trypanosomal, isoniazid, and antituberculosis. The first occurrence of each target word is bold. The references used in the text are among those identified as documents related to the words under analysis. It follows a specific logical pathway, although alternative interpretations may exist. Note: The text is prepared to represent the relationship between terms based on references. It does not provide a comprehensive analysis of the clinical efficacy of the mentioned drugs.

High-**throughput** screening (HTS) is a key method for drug discovery, enabling the simultaneous testing of large compound libraries. HTS is particularly valuable for identifying treatments for neglected diseases, as studies have shown (Annang et al. 2015; Williams et al. 2015; Kushwaha and Capalash 2022).

In **silico** approaches, based on computational methods, emerge as powerful complementary tools for discovering therapeutic compounds (Acevedo et al. 2017; Hassan et al. 2021; Porta et al. 2023). Notably, HTS benefits from integration with in silico tools (Andrade et al. 2019; Tavella et al. 2021). These computational strategies prove particularly effective for drug **repositioning**, identifying alternative therapeutic uses for existing compounds (Andrade et al. 2019).

Research highlights the importance of repurposing approaches for developing trypanosomiasis treatments, specifically identifying **eukaryotic** protein kinases as potential therapeutic targets against **trypanosomatid** diseases (Patel et al. 2013; Dichiara et al. 2017).

Kinases are not the only eukaryotic elements investigated as therapeutic targets in **trypanosomatids**. G-quadruplexes (G4s) are structures formed by nucleotides that play critical regulatory roles in humans. G4s are also crucial in protozoans, a finding that suggests their potential value for treating neglected diseases (Monti and Di Antonio 2023).

Additional repurposing examples with antitrypanosomal potential include **isatin**, originally developed as an anticancer drug (Leite et al. 2019; Chowdhary et al. 2022). Furthermore, human cyclic nucleotide phosphodiesterase (PDE) inhibitors, particularly PDE4 inhibitors, **display** potential for inhibiting **trypanosomal** phosphodiesterase B1 (TbrPDEB1) in Trypanosoma brucei.

Furoxanyl N-acylhydrazone derivatives also display therapeutic potential against trypanosomatids. Hybrid compounds containing imidazo[1,2-a]pyridine exhibit effective activity against Trypanosoma cruzi, while derivatives with 1,3-benzodioxol and 4-hydroxy-3-methoxybenzylidene show significant action against Leishmania amazonensis. Furthermore, evidence indicates that conjugating **isoniazid** (an **antituberculosis** drug) with furoxanyl N-acylhydrazone produces hybrid compounds with enhanced efficacy against multi-drug-resistant Mycobacterium tuberculosis strains, demonstrating this functional group's pharmacological potential for different situations, underscoring its pharmacological versatility within drug repositioning strategies (Hernández et al. 2013).

Box S2: Interaction with DeepSeek. This box documents an interaction with the DeepSeek web interface in April 2025.

Prompt: Write a paragraph relating these words: trypanosomatids, molecules, trypanosomatid, biologically, organisms, protozoan, chemical, chemotherapeutic, trypanosoma, parasites.

Response: Trypanosomatids are a family of biologically significant protozoan parasites, including the genus Trypanosoma, which cause diseases in humans and animals. These single-celled organisms rely on complex molecular interactions to survive and evade host immune responses. Researchers study trypanosomatid biology to identify key molecules that can be targeted by chemotherapeutic agents. Chemical interventions aim to disrupt essential pathways in these parasites, offering potential treatments for diseases like African sleeping sickness and Chagas disease. Understanding the unique biochemistry of trypanosomatids is crucial for developing effective drugs against these persistent and harmful organisms.

Prompt: Rewrite, ensuring that the term "molecules" appears specifically in the context of therapeutic molecules.

Response: Trypanosomatids, a family of protozoan parasites including Trypanosoma, are biologically complex organisms that cause devastating diseases. To combat these pathogens, researchers focus on identifying therapeutic molecules that disrupt essential parasite-specific processes. These chemical compounds serve as the foundation for novel chemotherapeutic agents, targeting trypanosomatid vulnerabilities while minimizing harm to the host. By studying the unique biology of these parasites, scientists aim to design molecules that inhibit critical pathways, offering hope for treating diseases like African sleeping sickness and Chagas disease. The development of such targeted molecules is key to overcoming the challenges posed by drug resistance in these persistent organisms.

Prompt: Add the context of drug repositioning.

Response: Trypanosomatids, a family of protozoan parasites including Trypanosoma, are biologically complex organisms responsible for neglected tropical diseases. Given the challenges in developing new drugs, researchers are increasingly exploring drug repositioning—repurposing existing therapeutic molecules originally designed for other conditions—as a faster, cost-effective strategy against these parasites. By screening known chemical compounds, scientists identify molecules with previously unrecognized anti-trypanosomatid activity, accelerating the development of chemotherapeutic options. This approach leverages existing safety and pharmacokinetic data, streamlining their use as potential treatments for diseases like African sleeping sickness and Chagas disease. Drug repositioning not only expands the arsenal against these pathogens but also offers a promising solution to combat drug resistance in these persistent organisms.

Table S1: Automated literature theme extraction. Thematic analysis generated from TXTree-exported vectors through unsupervised clustering and LLM processing (Qwen-1.8B). The table reflects the raw, unmodified output of the automated analysis.

#	Title	Analysis
0	Plant-derived triterpenes in neglected disease treatment	This cluster centers on the pharmacological properties of natural compounds, specifically triterpenes from a single plant species (Ochrosia elliptica), and their application in managing neglected tropical diseases. The focus lies on the unique biochemical profiles of these compounds and their potential therapeutic effects, emphasizing their role in addressing health disparities. The analysis highlights the interplay between plant-derived molecules, their structural characteristics, and their efficacy in treating diseases that are disproportionately affected by global health inequities. The methodology likely involves pharmacological studies, while the populations of interest are individuals with undiagnosed or underserved conditions.
1	Non-Coding RNA Mechanisms in Tropical Disease Pathogenesis	The cluster centers on the role of non-coding RNA in tropical diseases, focusing on pathogenesis and host-parasite interactions. It highlights the complex regulatory networks and molecular mechanisms that govern disease progression, with implications for therapeutic interventions targeting RNA-based pathways.
2	Essential Oil Components and Arbovirus Defense	This cluster examines the antimicrobial properties of essential oils and their application in combating arboviral infections. The analysis explores the chemical composition of plant-derived compounds and their potential to inhibit viral replication or modulate immune responses.
3	Hit Discovery and Drug Development	The cluster discusses the development of novel drug candidates, particularly hybrid compounds like furoxan hybrids, which target parasitic diseases. It emphasizes the integration of computational and experimental approaches to optimize drug safety and efficacy.
4	Vector-Parasite Roles	This cluster explores the ecological and biological roles of vectors like tabanids in transmitting parasitic diseases. The analysis highlights the complex interactions between vectors, pathogens, and host immune responses, with implications for vector control strategies.
5	Synthetic Compounds and Parasitic Evaluation	The cluster focuses on the synthesis of synthetic compounds and their evaluation for treating parasitic infections. It discusses the challenges of drug development for neglected diseases, including structural diversity and pharmacological selectivity.
6	Modeling, Spatial Statistics, Control	This cluster addresses the application of modeling and spatial statistics in understanding and controlling parasitic diseases. The analysis emphasizes the integration of computational methods with real-world data to inform public health strategies and resource allocation.
7	Targeted Drug Design, Transition States, Parasitic Pathogens	The cluster highlights the use of transition-state analogs and targeted drug design to combat parasitic infections. It discusses the importance of understanding molecular mechanisms to develop agents that disrupt critical metabolic pathways in parasites.
8	Natural Product Leads in Drug Discovery	This cluster explores the role of natural products in drug discovery for parasitic diseases, emphasizing the development of lead compounds from plant sources. The analysis highlights the potential of these compounds to address drug resistance and improve therapeutic outcomes.
9	Translational Success in Neglected Trypanosome Diseases	The cluster discusses the translation of research findings into clinical practice for treating trypanosome infections. It emphasizes the importance of overcoming barriers such as drug resistance and limited treatment options to improve patient outcomes.
10	Repurposing and Kinase Targeting	This cluster examines the repurposing of existing drugs, such as PDE4 inhibitors, for targeting kinases in parasitic diseases. The analysis highlights the potential of drug repositioning to address unmet medical needs in neglected tropical infections.

Continued on the next page

Table S1 – Continuation of the previous page

11	Vector-Specific Transmission in Savanna Ecosystems	The cluster focuses on the transmission dynamics of diseases caused by *Triatoma* vectors in savanna ecosystems. The analysis explores the role of vector behavior, environmental factors, and host susceptibility in disease spread, with implications for ecological and public health interventions.
12	Mitochondrial Targeting	This cluster addresses the development of drugs targeting mitochondrial function in trypanosomatids. The analysis highlights the integration of natural extracts and biochemical approaches to disrupt critical metabolic pathways and enhance drug efficacy.
13	Structural-Activity Relationships and Drug Development	The cluster discusses the use of structural-activity relationships (SARs) in drug discovery, emphasizing the development of compounds that optimize therapeutic properties. It highlights the importance of balancing molecular structure with pharmacological activity to address drug resistance.
14	AARS-Targeted Therapies	This cluster focuses on the molecular mechanisms of targeting aminoacyl-tRNA synthetases (AARS) as a therapeutic strategy for parasitic diseases. The analysis highlights the conserved roles of AARS in Leishmaniasis, African Trypanosomiasis, and Chagas disease, with implications for novel antimicrobial agents.
15	Nanosystems and Drug Discovery in Trypanosomatid Diseases	The cluster examines the integration of nanosystems, phytochemicals, and computational methods in drug discovery for parasitic diseases. It emphasizes the development of targeted therapies and the potential of nanopharmaceuticals to improve treatment outcomes for neglected tropical infections.

Supplementary Material References

Acevedo, Chonny Herrera et al. (Jan. 2017). "Computer-Aided Drug Design Using Sesquiterpene Lactones as Sources of New Structures with Potential Activity against Infectious Neglected Diseases". In: *Molecules (Basel, Switzerland)* 22.1. ISSN: 1420-3049. DOI: 10.3390/MOLECULES22010079. URL: https://doi.org/10.3390/MOLECULES22010079 (visited on 02/01/2026).

Andrade, Carolina Horta et al. (Mar. 2019). "In Silico Chemogenomics Drug Repositioning Strategies for Neglected Tropical Diseases". In: *Current medicinal chemistry* 26.23, pp. 4355–4379. ISSN: 1875-533X. DOI: 10.2174/0929867325666180309114824. URL: https://doi.org/10.2174/0929867325666180309114824 (visited on 02/01/2026).

Annang, F. et al. (Jan. 2015). "High-throughput screening platform for natural product-based drug discovery against 3 neglected tropical diseases: human African trypanosomiasis, leishmaniasis, and Chagas disease". In: *Journal of biomolecular screening* 20.1, pp. 82–91. ISSN: 1552-454X. DOI: 10.1177/1087057114555846. URL: https://doi.org/10.1177/1087057114555846 (visited on 02/01/2026).

Chowdhary, Shefali et al. (May 2022). "A Mini Review on Isatin, an Anticancer Scaffold with Potential Activities against Neglected Tropical Diseases (NTDs)". In: *Pharmaceuticals (Basel, Switzerland)* 15.5. ISSN: 1424-8247. DOI: 10.3390/PH15050536. URL: https://doi.org/10.3390/PH15050536 (visited on 02/01/2026).

Dichiara, Maria et al. (Aug. 2017). "Repurposing of Human Kinase Inhibitors in Neglected Protozoan Diseases". In: *ChemMedChem* 12.16, pp. 1235–1253. ISSN: 1860-7187. DOI: 10.1002/CMDC.201700259. URL: https://doi.org/10.1002/CMDC.201700259 (visited on 02/01/2026).

Hassan, Ahmed H.E. et al. (2021). "Pyrrolidine-based 3-deoxysphingosylphosphorylcholine analogs as possible candidates against neglected tropical diseases (NTDs): identification of hit compounds towards development of potential treatment of Leishmania donovani". In: *Journal of enzyme inhibition and medicinal chemistry* 36.1, pp. 1922–1930. ISSN: 1475-6374. DOI: 10.1080/14756366.2021.1969385. URL: https://doi.org/10.1080/14756366.2021.1969385 (visited on 02/01/2026).

Hernández, Paola et al. (Jan. 2013). "Hybrid furoxanyl N-acylhydrazone derivatives as hits for the development of neglected diseases drug candidates". In: *European journal of medicinal chemistry* 59, pp. 64–74. ISSN: 1768-3254. DOI: 10.1016/J.EJMECH.2012.10.047. URL: https://doi.org/10.1016/J.EJMECH.2012.10.047 (visited on 02/01/2026).

Kushwaha, Vikas and Neena Capalash (Sept. 2022). "Aminoacyl-tRNA synthetase (AARS) as an attractive drug target in neglected tropical trypanosomatid diseases-Leishmaniasis, Human African Trypanosomiasis and Chagas disease". In: *Molecular and biochemical parasitology* 251. ISSN: 1872-9428. DOI: 10.1016/J.MOLBIOPARA.2022.111510. URL: https://doi.org/10.1016/J.MOLBIOPARA.2022.111510 (visited on 02/01/2026).

Leite, Ana Cristina Lima et al. (Oct. 2019). "Privileged Structures in the Design of Potential Drug Candidates for Neglected Diseases". In: *Current medicinal chemistry* 26.23, pp. 4323–4354. ISSN: 1875-533X. DOI: 10.2174/0929867324666171023163752. URL: https://doi.org/10.2174/0929867324666171023163752 (visited on 02/01/2026).

Monti, Ludovica and Marco Di Antonio (June 2023). "G-Quadruplexes as Key Transcriptional Regulators in Neglected Trypanosomatid Parasites". In: *Chembiochem: a European journal of chemical biology* 24.12. ISSN: 1439-7633. DOI: 10.1002/CBIC.202300265. URL: https://doi.org/10.1002/CBIC.202300265 (visited on 02/01/2026).

Patel, Gautam et al. (May 2013). "Kinase scaffold repurposing for neglected disease drug discovery: discovery of an efficacious, lapatinib-derived lead compound for trypanosomiasis". In: *Journal of medicinal chemistry* 56.10, pp. 3820–3832. ISSN: 1520-4804. DOI: 10.1021/JM400349K. URL: https://doi.org/10.1021/JM400349K (visited on 02/01/2026).

Porta, Exequiel O.J. et al. (June 2023). "Systematic study of 1,2,3-triazolyl sterols for the development of new drugs against parasitic Neglected Tropical Diseases". In: *European journal of medicinal chemistry* 254. ISSN: 1768-3254. DOI: 10.1016/J.EJMECH.2023.115378. URL: https://doi.org/10.1016/J.EJMECH.2023.115378 (visited on 02/01/2026).

Tavella, T. A. et al. (Jan. 2021). "Yeast-based high-throughput screens for discovery of kinase inhibitors for neglected diseases". In: *Advances in protein chemistry and structural biology* 124, pp. 275–309. ISSN: 1876-1631. DOI: 10.1016/BS.APCSB.2020.09.007. URL: https://doi.org/10.1016/BS.APCSB.2020.09.007 (visited on 02/01/2026).

Williams, Kevin et al. (Mar. 2015). "Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases". In: *Journal of the Royal Society, Interface* 12.104. ISSN: 1742-5662. DOI: 10.1098/RSIF.2014.1289. URL: https://doi.org/10.1098/RSIF.2014.1289 (visited on 02/01/2026).

3 DISCUSSÃO

A combinação dos três artigos que compõem este estudo demonstra a aplicação e a posição do SWeePtex como alternativa à vetorização de textos. O SWeePtex representa uma transposição entre a Bioinformática e o processamento de linguagem natural, validando a premissa apresentada no Artigo 1 de que técnicas de representação de sequências biológicas podem ser aplicadas ao processamento de linguagem natural. O sucesso do método, ao gerar incorporações (*embeddings*) testadas qualitativamente e quantitativamente a partir de textos codificados em formato BSL, corrobora a analogia fundamental entre sequências biológicas e textuais, ambas caracterizadas por padrões sequenciais e dependências contextuais.

A Bioinformática é inspiração para o SWeePtex, ao passo que a base teórica é a projeção aleatória, fundamentada matematicamente pelo lema de Johnson-Lindenstrauss. Também alinhado ao paradigma da composição baseada em aleatoriedade, Kanerva (1994) explora a conexão entre métodos aleatórios e a cognição animal, aplicando o princípio na prática à representação textual por meio de indexação aleatória (Kanerva et al., 2000). A abordagem, no entanto, difere da projeção aleatória, uma vez que os conceitos são representados por vetores binários, com componentes originalmente atribuídos aleatoriamente. Apesar das diferenças metodológicas, ambas as técnicas compartilham a mesma fundamentação teórica, pois utilizam a aleatoriedade como ferramenta para compor uma base de representação distribuída e computacionalmente eficiente (Sahlgren, 2005).

O presente estudo demonstra que a projeção aleatória pode constituir uma alternativa viável ao paradigma dominante de aprendizagem profunda. A arquitetura do SWeePtex representa uma técnica de inteligência artificial que não se baseia na otimização por gradiente de milhões de parâmetros, mas sim em transformações geométricas interpretáveis. Esta abordagem encontra ressonância em perspectivas históricas sobre cognição e representação distribuída, nas quais a aleatoriedade atua como um mecanismo de abstração que preserva relações estruturais essenciais enquanto comprime a dimensionalidade representacional (Barlow; Rosenblith, 1961). O SWeePtex materializa esta perspectiva ao utilizar projeções aleatórias para mapear padrões textuais complexos em espaços vetoriais, oferecendo um contraponto teórico ao mecanismo de atenção dos *Transformers* (Vaswani et al., 2017).

Os vetores SWeePtex primários capturam inicialmente informações lexicais básicas de textos isolados, enquanto a representação semântica emerge por meio do processo de contextualização, mediado pelas médias vetoriais. O vetor contextual de cada palavra é calculado como a média dos vetores de todos os documentos em que aparece. Dessa forma, a palavra transcende sua representação isolada e passa

a incorporar o contexto semântico compartilhado entre os documentos em que ocorre, inspirado no princípio distribucional, de qual uma palavra é caracterizada pela companhia que mantém (Firth, 1957). Por sua vez, o vetor de cada documento é obtido pela média dos vetores contextuais de todas as palavras do documento. Este procedimento sintetiza semanticamente o documento a partir de seus termos já contextualizados, uma técnica consolidada para composição de representações de documentos (Mitchell; Lapata, 2010).

O SWeePtex apresenta um perfil de complexidade distinto do de modelos baseados em *Transformers*. Enquanto estes últimos exibem complexidade quadrática em relação ao comprimento da sequência devido aos mecanismos de atenção, o SWeePtex opera em tempo linear para a construção do vetor de alta dimensionalidade (HDV), seguido de uma projeção aleatória de custo fixo determinada pela dimensionalidade-alvo (LDV).

A aplicação bem-sucedida do SWeePtex em domínios distintos, ao longo dos três artigos (thioredoxin, DBpedia, doenças negligenciadas), além da exploração da literatura sobre Yoga já publicada (Ieger-Raittz et al., 2025), demonstra sua utilidade prática para a pesquisa científica. O processamento local em *hardware* doméstico representa uma forma de viabilizar análises profundas, reduzindo a dependência da infraestrutura de computação em nuvem. A possibilidade de executar análises completas em computadores pessoais torna metodologias sofisticadas de mineração de texto acessíveis a laboratórios e pesquisadores com recursos limitados.

O SWeeP gera um vetor de alta dimensionalidade (*High-Dimensional Vector*, HDV), que posteriormente é projetado em um espaço de dimensões reduzidas (*Lower-Dimensional Vector*, LDV) por meio de projeção aleatória. A dimensionalidade do HDV é determinada pelo número de caracteres distintos elevado à potência k, onde k é o tamanho da máscara, ou seja, o número de caracteres considerados simultaneamente na varredura (n^k). No contexto biológico típico, considerando a representação de 20 aminoácidos e uma máscara com $k = 3$ posições ativas, o HDV possui 8.000 *features* (20^3). Esse vetor pode então ser projetado em um espaço de dimensões arbitrárias, conforme a aplicação.

Ao codificar textos em formato BSL, o número de caracteres distintos diminui, permitindo que a representação utilize a mesma estrutura vetorial originalmente concebida para sequências biológicas. Retomando a analogia cognitiva, isso é semelhante à limitação perceptiva dos órgãos sensoriais, que facilita o processamento neurológico subsequente. Na prática, para uma máscara de tamanho $k = 3$, isso significa compactar o espaço vetorial de representação de 44 caracteres (26 letras, 10 dígitos, 6 sinais de pontuação, 1 espaço e 1 caractere genérico), de 85.184 dimensões (44^3), para 8.000 (20^3), reduzindo drasticamente o custo computacional. Com isso, o SWeePtex

opera com duas camadas de redução da complexidade: a primeira, com a codificação para BSL, que impacta todo o processo de geração por ser aplicada à composição inicial do HDV; e a segunda, com a projeção aleatória, que afeta o vetor final.

A capacidade de capturar padrões semânticos é corroborada pela alta correlação positiva observada entre as métricas do SWeePtex-Emb e as de modelos neurais, como verificado no Artigo 2. Essa correlação indica que ambos os métodos, ainda que com fundamentos distintos, conseguem identificar estruturas semelhantes nos dados textuais. Contudo, as avaliações revelam limitações específicas do SWeePtex. O método apresenta desempenho inferior em tarefas que demandam a compreensão de relações contextuais complexas, como a similaridade textual e a classificação em pares. Esta lacuna reflete a natureza da composição semântica por médias vetoriais, que não captura adequadamente relações de sinonímia contextual e de correferência não explícita; no entanto, é possível considerar alternativas para desenvolvimento futuro. A integração do SWeePtex-Emb com modelos leves, como redes rasas ou árvores de decisão, pode, possivelmente, introduzir capacidade inferencial sem renunciar ao baixo custo.

No aspecto qualitativo, em relação ao exemplo de uso do Artigo 3 e à publicação prévia (Ieger-Raittz et al., 2025), é notável que a abordagem permite a identificação de sinônimos, como demonstrado no caso das palavras “*leukemia*” e “*leukaemia*”. Ou seja, se houver relações léxicas entre os termos dos textos disponíveis no corpus, essas conexões são representadas.

Mais uma limitação é a insensibilidade à ordem das palavras na composição por média (Mitchell; Lapata, 2010). Contudo, este aspecto não é inerente ao paradigma da projeção aleatória, mas sim à estratégia específica de composição adotada. É concebível desenvolver técnicas alternativas com base na mesma fundamentação teórica. Por exemplo, em vez de computar apenas a média dos vetores dos documentos em que uma palavra ocorre, é possível calcular também a média dos vetores das palavras que aparecem em sequência a ela em todo o corpus, criando representações que capturam padrões sequenciais. Esta abordagem mantém a transparência computacional do método original, ao mesmo tempo em que incorpora informações de ordenação.

Um fator a considerar na avaliação do SWeePtex é a dificuldade intrínseca de realizar comparações justas com modelos de linguagem de larga escala. Enquanto os LLMs são tipicamente treinados em corpora massivos com infraestrutura computacional especializada, o treinamento do SWeePtex-Emb no Artigo 2 utiliza 630.000 documentos do DBpedia, processáveis em *hardware* pessoal. A diferença de escala torna impraticável a comparação direta da capacidade representacional. O SWeePtex não deve, portanto, ser avaliado como um competidor direto dos LLMs, mas sim como

uma possibilidade.

Em última análise, este estudo apresenta o SWeePtex como um promotor do paradigma da projeção aleatória, não como substituto dos LLMs, mas como alternativa capaz de impulsionar avanços práticos e teóricos no desenvolvimento científico e tecnológico. A discussão evidencia como a abordagem estimula a reflexão sobre o processo de vetorização, fundamental para compreender a lógica aritmética por trás da representação linguística e, consequentemente, cognitiva. Desse modo, favorece a compreensão epistemológica da construção de vetores linguísticos.

Por fim, o projeto torna o SWeePtex acessível programaticamente a partir do pacote Biotext (`https://pypi.org/p/biotext`). Também apresenta a aplicação em linha de comando TXTree (`https://sf.net/p/txtree`), que torna pública a abordagem de exploração de literatura via HTML-TM. Essas produções técnicas devem ser atualizadas ao longo de estudos futuros, assim como novas implementações associadas podem ser desenvolvidas. O estado aqui apresentado estabelece a base da linha de estudo sobre a incorporação de texto por meio de projeção aleatória inspirada em Bioinformática.

4 CONCLUSÃO

A utilização de técnicas baseadas em aleatoriedade na representação textual já possui documentação em trabalhos anteriores, mas a proposta aqui apresentada, inspirada na Bioinformática por meio do método SWeeP, constitui uma contribuição distinta e relevante para a área. A investigação apoia-se na compreensão teórica dos métodos de vetorização de texto, integrando preceitos matemáticos e conceitos das ciências cognitivas e linguísticas. A análise conduzida evidencia que a trajetória predominante do desenvolvimento tecnológico no processamento de textos é questionável sob as perspectivas epistemológicas e práticas.

A tendência atual concentra esforços na incorporação de informações por meio de abordagens de força bruta, focadas em grande escala, generalização e treinamento computacionalmente intensivo, frequentemente dependente de infraestrutura em nuvem. Esse paradigma tende a negligenciar a fundamentação teórica dos métodos, resultando em estruturas que demandam recursos exorbitantes e criam dependência tecnológica. Tais modelos dificultam a execução e a reprodutibilidade de experimentos, o que reforça a pertinência de paradigmas alternativos, como o da projeção aleatória.

Os experimentos conduzidos demonstram que a projeção aleatória por meio do SWeeP produz representações vetoriais linguísticas satisfatórias, com execução local em computador domesticamente acessível. O resultado prático, já validado pela publicação de uma revisão de literatura por pares, consolida a proposta central do estudo. Dessa forma, a proposta central do trabalho se valida. O legado metodológico da Bioinformática mostra-se transponível para a linguagem natural, o que se justifica pela natureza comum de ambas as áreas: o tratamento de estruturas sequenciais de elementos interdependentes.

Com este estudo, o escopo de aplicação do SWeeP se amplia, posicionando-o como uma ferramenta multifuncional cuja utilidade ultrapassa a aplicação original em sequências biológicas. Integrado ao Biotext sob a forma do SWeePtex, o método introduz a noção de camadas de simplificação funcional e teoricamente explicáveis, oferecendo, assim, uma contribuição substancial ao paradigma da projeção aleatória na representação textual e avançando a compreensão epistemológica sobre como a informação sequencial pode ser modelada.

REFERÊNCIAS

ABULAISH, M.; PARWEZ, M. A.; JAHIRUDDIN. **DiseaSE: A biomedical text analytics system for disease symptom extraction and characterization**. en. v. 100. [S.l.]: Elsevier BV, dez. 2019. P. 103324. DOI: 10.1016/j.jbi.2019.103324. Disponível em: https://doi.org/10.1016/j.jbi.2019.103324. Acesso em: 1 fev. 2026.

ACEVEDO, C. H.; SCOTTI, L.; ALVES, M. F.; DE FÁTIMA FORMIGA MELO DINIZ, M.; SCOTTI, M. T. Computer-Aided Drug Design Using Sesquiterpene Lactones as Sources of New Structures with Potential Activity against Infectious Neglected Diseases. **Molecules (Basel, Switzerland)**, Molecules, v. 22, n. 1, jan. 2017. ISSN 1420-3049. DOI: 10.3390/MOLECULES22010079. Disponível em: https://doi.org/10.3390/MOLECULES22010079. Acesso em: 1 fev. 2026.

ANDRADE, C. H.; NEVES, B. J.; MELO-FILHO, C. C.; RODRIGUES, J.; SILVA, D. C.; BRAGA, R. C.; CRAVO, P. V. L. In Silico Chemogenomics Drug Repositioning Strategies for Neglected Tropical Diseases. **Current medicinal chemistry**, Curr Med Chem, v. 26, n. 23, p. 4355–4379, mar. 2019. ISSN 1875-533X. DOI: 10.2174/0929867325666180309114824. Disponível em: https://doi.org/10.2174/0929867325666180309114824. Acesso em: 1 fev. 2026.

ANNANG, F.; PÉREZ-MORENO, G.; GARCÍA-HERNÁNDEZ, R.; CORDON-OBRAS, C.; MARTÍN, J.; TORMO, J. R.; RODRÍGUEZ, L.; DE PEDRO, N.; GÓMEZ-PÉREZ, V.; VALENTE, M.; REYES, F.; GENILLOUD, O.; VICENTE, F.; CASTANYS, S.; RUIZ-PÉREZ, L. M.; NAVARRO, M.; GAMARRO, F.; GONZÁLEZ-PACANOWSKA, D. High-throughput screening platform for natural product-based drug discovery against 3 neglected tropical diseases: human African trypanosomiasis, leishmaniasis, and Chagas disease. **Journal of biomolecular screening**, J Biomol Screen, v. 20, n. 1, p. 82–91, jan. 2015. ISSN 1552-454X. DOI: 10.1177/1087057114555846. Disponível em: https://doi.org/10.1177/1087057114555846. Acesso em: 1 fev. 2026.

ARAUJO, J. D.; SANTOS-E-SILVA, J. C.; COSTA-MARTINS, A. G.; SAMPAIO, V.; CASTRO, D. B. de; SOUZA, R. F. de; GIDDALURU, J.; RAMOS, P. I. P.; PITA, R.; BARRETO, M. L.; BARRAL-NETTO, M.; NAKAYA, H. I. Tucuxi-BLAST: Enabling fast and accurate record linkage of large-scale health-related administrative databases through a DNA-encoded approach. **PeerJ**, PeerJ Inc., v. 10, e13507, jul. 2022. ISSN

2167-8359. DOI: 10.7717/peerj.13507. Disponível em: https://doi.org/10.7717/peerj.13507. Acesso em: 1 fev. 2026.

ASGARI, E.; MOFRAD, M. R. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. **PLOS ONE**, Public Library of Science, v. 10, n. 11, e0141287, nov. 2015. ISSN 1932-6203. DOI: 10.1371/JOURNAL.PONE.0141287. Disponível em: https://doi.org/10.1371/JOURNAL.PONE.0141287. Acesso em: 1 fev. 2026.

AZIZ, M.; POPA, I.; ZIA, A.; FISCHER, A.; KHAN, S. A.; HAMEDANI, A. F.; ASIF, A. R. **KnowVID-19: A knowledge-based system to extract targeted COVID-19 information from online medical repositories**. en. v. 14. [S.l.]: MDPI AG, nov. 2024. P. 1411. DOI: 10.3390/biom14111411. Disponível em: https://doi.org/10.3390/biom14111411. Acesso em: 1 fev. 2026.

BARLOW, H. B.; ROSENBLITH, W. A. Possible principles underlying the transformations of sensory messages. In: SENSORY Communication. [S.l.]: MIT Press, 1961. P. 217–234. Disponível em: https://www.cnbc.cmu.edu/~tai/microns_papers/Barlow-SensoryCommunication-1961.pdf. Acesso em: 1 fev. 2026.

BELTAGY, I.; PETERS, M. E.; COHAN, A. Longformer: The Long-Document Transformer. arXiv, 2020. DOI: 10.48550/arXiv.2004.05150. Disponível em: http://doi.org/10.48550/arXiv.2004.05150. Acesso em: 1 fev. 2026.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JANVIN, C. A Neural Probabilistic Language Model. **Journal of Machine Learning Research**, v. 3, p. 1137–1155, 2003. DOI: 10.5555/944919.944966. Disponível em: https://dl.acm.org/doi/abs/10.5555/944919.944966. Acesso em: 1 fev. 2026.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. Sebastopol, CA: O'Reilly Media, jul. 2009.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. **Enriching Word Vectors with Subword Information**. [S.l.: s.n.], 2016. DOI: 10.48550/arXiv.1607.04606. Disponível em: https://doi.org/10.48550/arXiv.1607.04606. Acesso em: 1 fev. 2026.

CANESE, K.; WEIS, S. **PubMed: The Bibliographic Database**. [S.l.: s.n.], 2002. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK153385/pdf/Bookshelf_NBK153385.pdf. Acesso em: 1 fev. 2026.

CHEN, Q.; AI, N.; LIAO, J.; SHAO, X.; LIU, Y.; FAN, X. **Revealing topics and their evolution in biomedical literature using Bio-DTM: a case study of ginseng**. en. v. 12. [S.l.]: Springer Science e Business Media LLC, set. 2017. P. 27. DOI: 10.1186/s13020-017-0148-7. Disponível em: https://doi.org/10.1186/s13020-017-0148-7. Acesso em: 1 fev. 2026.

CHILD, R.; GRAY, S.; RADFORD, A.; SUTSKEVER, I. Generating Long Sequences with Sparse Transformers. arXiv, 2019. DOI: 10.48550/arXiv.1904.10509. Disponível em: http://doi.org/10.48550/arXiv.1904.10509. Acesso em: 1 fev. 2026.

CHOWDHARY, S.; SHALINI; ARORA, A.; KUMAR, V. A Mini Review on Isatin, an Anticancer Scaffold with Potential Activities against Neglected Tropical Diseases (NTDs). **Pharmaceuticals (Basel, Switzerland)**, Pharmaceuticals (Basel), v. 15, n. 5, mai. 2022. ISSN 1424-8247. DOI: 10.3390/PH15050536. Disponível em: https://doi.org/10.3390/PH15050536. Acesso em: 1 fev. 2026.

COCK, P. J. A.; ANTAO, T.; CHANG, J. T.; CHAPMAN, B. A.; COX, C. J.; DALKE, A.; FRIEDBERG, I.; HAMELRYCK, T.; KAUFF, F.; WILCZYNSKI, B.; DE HOON, M. J. L. **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. v. 25. [S.l.: s.n.], 2009. P. 1422–1423. DOI: 10.1093/bioinformatics/btp163. Disponível em: https://doi.org/10.1093/bioinformatics/btp163. Acesso em: 1 fev. 2026.

CURRIE, G. M. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? **Seminars in Nuclear Medicine**, v. 53, n. 5, p. 719–730, 2023. DOI: 10.1053/j.semnuclmed.2023.04.008. Disponível em: https://doi.org/10.1053/j.semnuclmed.2023.04.008. Acesso em: 1 fev. 2026.

DAI, H.-J.; SU, C.-H.; LAI, P.-T.; HUANG, M.-S.; JONNAGADDALA, J.; ROSE JUE, T.; RAO, S.; CHOU, H.-J.; MILACIC, M.; SINGH, O.; SYED-ABDUL, S.; HSU, W.-L. **MET network in PubMed: a text-mined network visualization and curation system**. en. v. 2016. [S.l.]: Oxford University Press (OUP), mai. 2016. baw090. DOI: 10.1093/database/baw090. Disponível em: https://doi.org/10.1093/database/baw090. Acesso em: 1 fev. 2026.

DE PIERRI, C. R.; VOYCEIK, R.; SANTOS DE MATTOS, L. G. C.; KULIK, M. G.; CAMARGO, J. O.; REPULA DE OLIVEIRA, A. M.; LIMA NICHIO, B. T. de; MARCHAUKOSKI, J. N.; SILVA FILHO, A. C. da; GUIZELINI, D.; ORTEGA, J. M.; PEDROSA, F. O.; RAITTZ, R. T. SWeeP: representing large biological sequences datasets in compact vectors. **Scientific Reports**, Nature Research, v. 10, n. 1, p. 91, jan. 2020. ISSN 2045-2322. DOI: 10.1038/s41598-019-55627-4. Disponível em: https://doi.org/10.1038/s41598-019-55627-4. Acesso em: 1 fev. 2026.

DEEPSEEK-AI et al. **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**. [S.l.: s.n.], 2025. DOI: 10.48550/arXiv.2501.12948. Disponível em: https://doi.org/10.48550/arXiv.2501.12948. Acesso em: 1 fev. 2026.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv**, 2018. DOI: 10.48550/arXiv.1810.04805. Disponível em: https://doi.org/10.48550/arXiv.1810.04805. Acesso em: 1 fev. 2026.

DICHIARA, M.; MARRAZZO, A.; PREZZAVENTO, O.; COLLINA, S.; RESCIFINA, A.; AMATA, E. Repurposing of Human Kinase Inhibitors in Neglected Protozoan Diseases. **ChemMedChem**, ChemMedChem, v. 12, n. 16, p. 1235–1253, ago. 2017. ISSN 1860-7187. DOI: 10.1002/CMDC.201700259. Disponível em: https://doi.org/10.1002/CMDC.201700259. Acesso em: 1 fev. 2026.

ELMAN, J. L. Finding Structure in Time. **Cognitive Science**, v. 14, n. 2, p. 179–211, 1990. DOI: 10.1207/s15516709cog1402_1. Disponível em: https://doi.org/10.1207/s15516709cog1402_1. Acesso em: 1 fev. 2026.

FIRTH, J. R. **Papers in Linguistics 1934-1951**. London: Oxford University Press, 1957. Disponível em: https://archive.org/details/papersinlinguist0000firt. Acesso em: 1 fev. 2026.

FLYAMER, I. et al. Phlya/adjustText: 1.3.0. **Zenodo**, Zenodo, out. 2024. DOI: 10.5281/zenodo.14019059. Disponível em: https://doi.org/10.5281/zenodo.14019059. Acesso em: 1 fev. 2026.

GANGULI, S.; SOMPOLINSKY, H. Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis. **Annual Review of Neuroscience**, Annual Reviews, v. 35, n. 1, p. 485–508, jul. 2012. ISSN

1545-4126. DOI: 10.1146/annurev-neuro-062111-150410. Disponível em: https://doi.org/10.1146/annurev-neuro-062111-150410. Acesso em: 1 fev. 2026.

GOBEILL, J.; CAUCHETEUR, D.; MICHEL, P.-A.; MOTTIN, L.; PASCHE, E.; RUCH, P. **SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts**. en. v. 48. [S.l.]: Oxford University Press (OUP), jul. 2020. w12–w16. DOI: 10.1093/nar/gkaa328. Disponível em: https://doi.org/10.1093/nar/gkaa328. Acesso em: 1 fev. 2026.

GRIFFITH, O. L.; MONTGOMERY, S. B.; BRIDGET, B.; CHU, B.; KASAIAN, K.; AERTS, S.; MAHONY, S.; SLEUMER, M. C.; BILENKY, M.; HAEUSSLER, M.; GRIFFITH, M.; M, G. S.; GIARDINE, B.; HOOGHE, B.; VAN LOO, P.; BLANCO, E.; TICOLL, A.; LITHWICK, S.; PORTALES-CASAMAR, E.; DONALDSON, I. J.; ROBERTSON, G.; WADELIUS, C.; DE BLESER, P.; VLIEGHE, D.; HALFON, M. S.; WASSERMAN, W.; HARDISON, R.; BERGMAN, C. M.; JONES, S. J. M.; CONSORTIUM, O. R. A. **ORegAnno: an open-access community-driven resource for regulatory annotation**. en. v. 36. [S.l.]: Oxford University Press (OUP), jan. 2008. P. d107–13. DOI: 10.1093/nar/gkm967. Disponível em: https://doi.org/10.1093/nar/gkm967. Acesso em: 1 fev. 2026.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; KERKWIJK, M. H. van; BRETT, M.; HALDANE, A.; RÍO, J. F. del; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. Array programming with NumPy. **Nature 2020 585:7825**, Nature Publishing Group, v. 585, n. 7825, p. 357–362, set. 2020. ISSN 1476-4687. DOI: 10.1038/s41586-020-2649-2. Disponível em: https://doi.org/10.1038/s41586-020-2649-2. Acesso em: 1 fev. 2026.

HASSAN, A. H.; PHAN, T. N.; YOON, S.; LEE, C. J.; JEON, H. R.; KIM, S. H.; NO, J. H.; LEE, Y. S. Pyrrolidine-based 3-deoxysphingosylphosphorylcholine analogs as possible candidates against neglected tropical diseases (NTDs): identification of hit compounds towards development of potential treatment of Leishmania donovani. **Journal of enzyme inhibition and medicinal chemistry**, J Enzyme Inhib Med Chem, v. 36, n. 1, p. 1922–1930, 2021. ISSN 1475-6374. DOI: 10.1080/14756366.2021.1969385. Disponível em: https://doi.org/10.1080/14756366.2021.1969385. Acesso em: 1 fev. 2026.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M. T.; YEGANEGI, M. R. Text Mining in Big Data Analytics. **Big Data and Cognitive Computing**, Multidisciplinary Digital Publishing Institute, v. 4, n. 1, p. 1, jan. 2020. ISSN 2504-2289. DOI: 10.3390/bdcc4010001. Disponível em: https://doi.org/10.3390/bdcc4010001. Acesso em: 1 fev. 2026.

HAYEN, K.; CONTRIBUTORS, N. **Nuitka the Python Compiler**. [S.l.]: Nuitka Project, 2025. Disponível em: https://nuitka.net. Acesso em: 1 fev. 2026.

HERNÁNDEZ, P.; ROJAS, R.; GILMAN, R. H.; SAUVAIN, M.; LIMA, L. M.; BARREIRO, E. J.; GONZÁLEZ, M.; CERECETTO, H. Hybrid furoxanyl N-acylhydrazone derivatives as hits for the development of neglected diseases drug candidates. **European journal of medicinal chemistry**, Eur J Med Chem, v. 59, p. 64–74, jan. 2013. ISSN 1768-3254. DOI: 10.1016/J.EJMECH.2012.10.047. Disponível em: https://doi.org/10.1016/J.EJMECH.2012.10.047. Acesso em: 1 fev. 2026.

HOFFMANN, M. F.; PREISSNER, S. C.; JANETTE, N.; DUNKEL, M.; PREISSNER, R.; PREISSNER, S. **The Transformer database: biotransformation of xenobiotics**. en. v. 42. [S.l.]: Oxford University Press (OUP), jan. 2014. P. d1113–7. DOI: 10.1093/nar/gkt1246. Disponível em: https://doi.org/10.1093/nar/gkt1246. Acesso em: 1 fev. 2026.

HOWARD, B. E.; PHILLIPS, J.; MILLER, K.; TANDON, A.; MAV, D.; SHAH, M. R.; STEPHANIE, H.; PELCH, K. E.; WALKER, V.; A, R. A.; MACLEOD, M.; SHAH, R. R.; THAYER, K. **SWIFT-Review: a text-mining workbench for systematic review**. en. v. 5. [S.l.]: Springer Nature, mai. 2016. P. 87. DOI: 10.1186/s13643-016-0263-z. Disponível em: https://doi.org/10.1186/s13643-016-0263-z. Acesso em: 1 fev. 2026.

HUANG, H.; ARIGHI, C. N.; ROSS, K. E.; REN, J.; LI, G.; CHEN, S.-C.; WANG, Q.; COWART, J.; VIJAY-SHANKER, K.; WU, C. H. **iPTMnet: an integrated resource for protein post-translational modification network discovery**. v. 46. [S.l.]: Oxford University Press (OUP), jan. 2018. P. d542–d550. DOI: 10.1093/nar/gkx1104. Disponível em: https://doi.org/10.1093/nar/gkx1104. Acesso em: 1 fev. 2026.

HUANG, Y.; WANG, L.; ZAN, L.-S. **ARN: Analysis and visualization system for adipogenic regulation network information**. en. v. 6. [S.l.]: Springer Science e

Business Media LLC, dez. 2016. DOI: 10.1038/srep39347. Disponível em: https://doi.org/10.1038/srep39347. Acesso em: 1 fev. 2026.

HUERTA-CEPAS, J.; SERRA, F.; BORK, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. **Molecular Biology and Evolution**, v. 33, p. 1635–1638, 2016. DOI: 10.1093/molbev/msw046. Disponível em: https://doi.org/10.1093/molbev/msw046. Acesso em: 1 fev. 2026.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, mai. 2007. ISSN 1521-9615. DOI: 10.1109/MCSE.2007.55. Disponível em: https://doi.org/10.1109/MCSE.2007.55. Acesso em: 1 fev. 2026.

HUTSON, M. Forget ChatGPT: why researchers now run small AIs on their laptops. **Nature**, v. 633, n. 8030, p. 728–729, set. 2024. ISSN 0028-0836. DOI: 10.1038/D41586-024-02998-Y. Disponível em: https://doi.org/10.1038/D41586-024-02998-Y. Acesso em: 1 fev. 2026.

IEGER-RAITTZ, R.; DE PIERRI, C. R.; PERICO, C. P.; FATIMA COSTA, F. de; BANA, E. G.; VICENZI, L.; JESUS SOARES MACHADO, D. de; MARCHAUKOSKI, J. N.; RAITTZ, R. T. What are we learning with Yoga? Mapping the scientific literature on Yoga using a vector-text-mining approach. **PLOS ONE**, Public Library of Science, v. 20, n. 5, e0322791, mai. 2025. ISSN 1932-6203. DOI: 10.1371/JOURNAL.PONE.0322791. Disponível em: https://doi.org/10.1371/JOURNAL.PONE.0322791. Acesso em: 1 fev. 2026.

IVANISENKO, T. V.; SAIK, O. V.; DEMENKOV, P. S.; IVANISENKO, N. V.; SAVOSTIANOV, A. N.; IVANISENKO, V. A. **ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature**. en. v. 21. [S.l.]: Springer Science e Business Media LLC, set. 2020. P. 228. DOI: 10.1186/s12859-020-03557-8. Disponível em: https://doi.org/10.1186/s12859-020-03557-8. Acesso em: 1 fev. 2026.

JANG, H.; LIM, J.; LIM, J.-H.; PARK, S.-J.; LEE, K.-C. **BioProber: Software system for biomedical relation discovery from PubMed**. [S.l.]: IEEE, ago. 2006. DOI: 10.1109/IEMBS.2006.259838. Disponível em: https://doi.org/10.1109/IEMBS.2006.259838. Acesso em: 1 fev. 2026.

JAYLET, T.; COUSTILLET, T.; JORNOD, F.; MARGARITTE-JEANNIN, P.; AUDOUZE, K. **AOP-helpFinder 2.0: Integration of an event-event searches module**. en. v. 177. [S.l.]: Elsevier BV, jul. 2023. P. 108017. DOI: 10.1016/j.envint.2023.108017. Disponível em: https://doi.org/10.1016/j.envint.2023.108017. Acesso em: 1 fev. 2026.

JOHNSON, W. B.; LINDENSTRAUSS, J. Extensions of Lipschitz mappings into a Hilbert space. In: CONTEMPORARY Mathematics. [S.l.]: American Mathematical Society, 1984. P. 189–206. DOI: 10.1090/conm/026/737400. Disponível em: https://doi.org/10.1090/conm/026/737400. Acesso em: 1 fev. 2026.

JONES, S. K. **A Statistical Interpretation of Term Specificity and its Application in Retrieval**. v. 28. [S.l.: s.n.], 1972. P. 11–21. Disponível em: https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf. Acesso em: 1 fev. 2026.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd. [S.l.]: Stanford University, 2025. Disponível em: https://web.stanford.edu/~jurafsky/slp3. Acesso em: 1 fev. 2026.

KAFKAS,.; DUNHAM, I.; MCENTYRE, J. **Literature evidence in open targets - a target validation platform**. v. 8. [S.l.]: Springer Nature, dez. 2017. DOI: 10.1186/s13326-017-0131-3. Disponível em: https://doi.org/10.1186/s13326-017-0131-3. Acesso em: 1 fev. 2026.

KANERVA, P. The Spatter Code for Encoding Concepts at Many Levels. In: ICANN 94. London: Springer London, 1994. P. 226–229. DOI: 10.1007/978-1-4471-2097-1_52. Disponível em: https://doi.org/10.1007/978-1-4471-2097-1_52. Acesso em: 1 fev. 2026.

KANERVA, P.; KRISTOFERSON, J.; HOLS, A. **Random Indexing of Text Samples for Latent Semantic Analysis**. [S.l.: s.n.], 2000. Disponível em: https://escholarship.org/uc/item/5644k0w6. Acesso em: 1 fev. 2026.

KASKI, S. Dimensionality reduction by random mapping: fast similarity computation for clustering. In: 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227). [S.l.: s.n.], 1998. P. 413–418. DOI: 10.1109/IJCNN.1998.682302.

Disponível em: https://doi.org/10.1109/IJCNN.1998.682302. Acesso em: 1 fev. 2026.

KASKI, S.; HONKELA, T.; LAGUS, K.; KOHONEN, T. WEBSOM Self-organizing maps of document collections. **Neurocomputing**, v. 21, n. 1, p. 101–117, 1998. ISSN 0925-2312. DOI: 10.1016/S0925-2312(98)00039-3. Disponível em: https://doi.org/10.1016/S0925-2312(98)00039-3. Acesso em: 1 fev. 2026.

KASNECI, E.; SESSLER, K.; KÜCHEMANN, S.; BANNERT, M.; DEMENTIEVA, D.; FISCHER, F.; GASSER, U.; GROH, G.; GÜNNEMANN, S.; HÜLLERMEIER, E.; KRUSCHE, S.; KUTYNIOK, G.; MICHAELI, T.; NERDEL, C.; PFEFFER, J.; POQUET, O.; SAILER, M.; SCHMIDT, A.; SEIDEL, T.; STADLER, M.; WELLER, J.; KUHN, J.; KASNECI, G. ChatGPT for good? On opportunities and challenges of large language models for education. **Learning and Individual Differences**, JAI, v. 103, p. 102274, abr. 2023. ISSN 1041-6080. DOI: 10.1016/J.LINDIF.2023.102274. Disponível em: https://doi.org/10.1016/J.LINDIF.2023.102274. Acesso em: 1 fev. 2026.

KUO, T.-C.; TIAN, T.-F.; TSENG, Y. J. **3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data**. en. v. 7. [S.l.]: Springer Science e Business Media LLC, jul. 2013. P. 64. DOI: 10.1186/1752-0509-7-64. Disponível em: https://doi.org/10.1186/1752-0509-7-64. Acesso em: 1 fev. 2026.

KURATOV, Y.; ARKHIPOV, M. **Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language**. [S.l.: s.n.], 2019. DOI: 10.48550/arXiv.1905.07213. Disponível em: https://doi.org/10.48550/arXiv.1905.07213. Acesso em: 1 fev. 2026.

KUSHWAHA, V.; CAPALASH, N. Aminoacyl-tRNA synthetase (AARS) as an attractive drug target in neglected tropical trypanosomatid diseases-Leishmaniasis, Human African Trypanosomiasis and Chagas disease. **Molecular and biochemical parasitology**, Mol Biochem Parasitol, v. 251, set. 2022. ISSN 1872-9428. DOI: 10.1016/J.MOLBIOPARA.2022.111510. Disponível em: https://doi.org/10.1016/J.MOLBIOPARA.2022.111510. Acesso em: 1 fev. 2026.

LEHMANN, J.; ISELE, R.; JAKOB, M.; JENTZSCH, A.; KONTOKOSTAS, D.; MENDES, P. N.; HELLMANN, S.; MORSEY, M.; KLEEF, P. van; AUER, S.; BIZER, C. DBpedia A large-scale, multilingual knowledge base extracted from Wikipedia.

Semantic Web, v. 6, n. 2, p. 167–195, 2015. DOI: 10.3233/SW-140134. Disponível em: https://doi.org/10.3233/SW-140134. Acesso em: 1 fev. 2026.

LEIMEISTER, C. A.; SCHELLHORN, J.; DÖRRER, S.; GERTH, M.; BLEIDORN, C.; MORGENSTERN, B. Prot-SpaM: fast alignment-free phylogeny reconstruction based on whole-proteome sequences. **GigaScience**, Oxford Academic, v. 8, n. 3, p. 1–14, mar. 2019. ISSN 2047217X. DOI: 10.1093/GIGASCIENCE/GIY148. Disponível em: https://doi.org/10.1093/GIGASCIENCE/GIY148. Acesso em: 1 fev. 2026.

LEITE, A. C. L.; ESPÍNDOLA, J. W. P.; OLIVEIRA CARDOSO, M. V. de; OLIVEIRA FILHO, G. B. de. Privileged Structures in the Design of Potential Drug Candidates for Neglected Diseases. **Current medicinal chemistry**, Curr Med Chem, v. 26, n. 23, p. 4323–4354, out. 2019. ISSN 1875-533X. DOI: 10.2174/0929867324666171023163752. Disponível em: https://doi.org/10.2174/0929867324666171023163752. Acesso em: 1 fev. 2026.

LI, J.; GAO, J.; FENG, B.; JING, Y. **PlagueKD: a knowledge graph-based plague knowledge database**. en. v. 2022. [S.l.]: Oxford University Press (OUP), nov. 2022. DOI: 10.1093/database/baac100. Disponível em: https://doi.org/10.1093/database/baac100. Acesso em: 1 fev. 2026.

LI, S.; HU, R.; WANG, L. **Efficiently Building a Domain-Specific Large Language Model from Scratch: A Case Study of a Classical Chinese Large Language Model**. [S.l.: s.n.], 2025. DOI: 10.48550/arXiv.2505.11810. Disponível em: https://doi.org/10.48550/arXiv.2505.11810. Acesso em: 1 fev. 2026.

LILLEBERG, J.; ZHU, Y.; ZHANG, Y. Support vector machines and Word2vec for text classification with semantic features. **Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2015**, Institute of Electrical e Electronics Engineers Inc., p. 136–140, set. 2015. DOI: 10.1109/ICCI-CC.2015.7259377. Disponível em: https://doi.org/10.1109/ICCI-CC.2015.7259377. Acesso em: 1 fev. 2026.

MA, L.; ZHANG, Y. Using Word2Vec to process big text data. **Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015**, Institute of Electrical e Electronics Engineers Inc., p. 2895–2897, dez. 2015. DOI: 10.1109/BIGDATA.2015.7364114. Disponível em: https://doi.org/10.1109/BIGDATA.2015.7364114. Acesso em: 1 fev. 2026.

MACHADO, D. d. J. S.; DE PIERRI, C. R.; SANTOS, L. G. C.; SCAPIN, L.; SILVA FILHO, A. C. da; PERICO, C. P.; PEDROSA, F. d. O.; RAITTZ, R. T. Biotext: Exploiting Biological-Text Format for Text Mining. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2021.04.08.439078, abr. 2021. DOI: 10.1101/2021.04.08.439078. Disponível em: https://doi.org/10.1101/2021.04.08.439078. Acesso em: 1 fev. 2026.

MACNEE, M.; PÉREZ-PALMA, E.; SCHUMACHER-BASS, S.; DALTON, J.; LEU, C.; DANIEL, B.; LAL, D. **SimText: a text mining framework for interactive analysis and visualization of similarities among biomedical entities**. en. v. 37. [S.l.]: Oxford University Press (OUP), nov. 2021. P. 4285–4287. DOI: 10.1093/bioinformatics/btab365. Disponível em: https://doi.org/10.1093/bioinformatics/btab365. Acesso em: 1 fev. 2026.

MCKINNEY, W. Data Structures for Statistical Computing in Python. In: PROCEEDINGS of the 9th Python in Science Conference. [S.l.: s.n.], 2010. P. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

METZGER, V. T.; CANNON, D. C.; YANG, J. J.; MATHIAS, S. L.; BOLOGA, C. G.; WALLER, A.; SCHÜRER, S. C.; VIDOVI, D.; KELLEHER, K. J.; SHEILS, T. K.; JUHL, J. L.; LAMBERT, C. G.; OPREA, T. I.; EDWARDS, J. S. **TIN-X version 3: update with expanded dataset and modernized architecture for enhanced illumination of understudied targets**. en. v. 12. [S.l.]: PeerJ, jun. 2024. e17470. DOI: 10.7717/peerj.17470. Disponível em: https://doi.org/10.7717/peerj.17470. Acesso em: 1 fev. 2026.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. **arXiv**, jan. 2013. DOI: 10.48550/arXiv.1301.3781. Disponível em: https://doi.org/10.48550/arXiv.1301.3781. Acesso em: 1 fev. 2026.

MITCHELL, J.; LAPATA, M. Composition in Distributional Models of Semantics. **Cognitive Science**, John Wiley & Sons, Ltd, v. 34, n. 8, p. 1388–1429, nov. 2010. ISSN 1551-6709. DOI: 10.1111/J.1551-6709.2010.01106.X. Disponível em: https://doi.org/10.1111/J.1551-6709.2010.01106.X. Acesso em: 1 fev. 2026.

MONTI, L.; DI ANTONIO, M. G-Quadruplexes as Key Transcriptional Regulators in Neglected Trypanosomatid Parasites. **Chembiochem: a European journal of chemical biology**, Chembiochem, v. 24, n. 12, jun. 2023. ISSN 1439-7633. DOI:

10.1002/CBIC.202300265. Disponível em: https://doi.org/10.1002/CBIC.202300265. Acesso em: 1 fev. 2026.

MUELLER, A. C. **Wordcloud**. [S.l.: s.n.], 2026. Disponível em: https://github.com/amueller/word_cloud. Acesso em: 1 fev. 2026.

MUENNIGHOFF, N.; TAZI, N.; MAGNE, L.; REIMERS, N. MTEB: Massive Text Embedding Benchmark. **EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference**, Association for Computational Linguistics (ACL), p. 2006–2029, out. 2022. DOI: 10.18653/v1/2023.eacl-main.148. Disponível em: https://aclanthology.org/2023.eacl-main.148. Acesso em: 1 fev. 2026.

NEVES, M. **Integration of the PubAnnotation ecosystem in the development of a web-based search tool for alternative methods**. en. v. 18. [S.l.]: Korea Genome Organization, jun. 2020. e18. DOI: 10.5808/GI.2020.18.2.e18. Disponível em: https://doi.org/10.5808/GI.2020.18.2.e18. Acesso em: 1 fev. 2026.

NOVÁEK, V.; BURNS, G. A. SKIMMR: facilitating knowledge discovery in life sciences by machine-aided skim reading. **PeerJ**, PeerJ, v. 2, e483, jul. 2014. ISSN 2167-8359. DOI: 10.7717/peerj.483. Disponível em: http://doi.org/10.7717/peerj.483. Acesso em: 1 fev. 2026.

OLLAMA. **Ollama's documentation**. [S.l.]: Ollama, 2026. Disponível em: https://docs.ollama.com. Acesso em: 1 fev. 2026.

OPENAI et al. GPT-4 Technical Report. **arXiv**, 2023. DOI: 10.48550/arXiv.2303.08774. Disponível em: https://doi.org/10.48550/arXiv.2303.08774. Acesso em: 1 fev. 2026.

PATEL, G.; KARVER, C. E.; BEHERA, R.; GUYETT, P. J.; SULLENBERGER, C.; EDWARDS, P.; RONCAL, N. E.; MENSA-WILMOT, K.; POLLASTRI, M. P. Kinase scaffold repurposing for neglected disease drug discovery: discovery of an efficacious, lapatinib-derived lead compound for trypanosomiasis. **Journal of medicinal chemistry**, J Med Chem, v. 56, n. 10, p. 3820–3832, mai. 2013. ISSN 1520-4804. DOI: 10.1021/JM400349K. Disponível em: https://doi.org/10.1021/JM400349K. Acesso em: 1 fev. 2026.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; M., B.; PERROT, M.; DUCHESNAY, E. **Scikit-learn: Machine Learning in Python**. v. 12. [S.l.: s.n.], 2011. P. 2825–2830. Disponível em: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Acesso em: 1 fev. 2026.

PERICO, C. P.; PIERRI, C. R. D.; NETO, G. P.; FERNANDES, D. R.; PEDROSA, F. O.; SOUZA, E. M. de; RAITTZ, R. T. Genomic landscape of the SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin. **Frontiers in Microbiology**, v. 13, 2022. DOI: 10.3389/fmicb.2022.1037455. Disponível em: https://doi.org/10.3389/fmicb.2022.1037455. Acesso em: 1 fev. 2026.

PORTA, E. O.; BALLARI, M. S.; CARLUCCI, R.; WILKINSON, S.; MA, G.; TEKWANI, B. L.; LABADIE, G. R. Systematic study of 1,2,3-triazolyl sterols for the development of new drugs against parasitic Neglected Tropical Diseases. **European journal of medicinal chemistry**, Eur J Med Chem, v. 254, jun. 2023. ISSN 1768-3254. DOI: 10.1016/J.EJMECH.2023.115378. Disponível em: https://doi.org/10.1016/J.EJMECH.2023.115378. Acesso em: 1 fev. 2026.

RADFORD, A.; NARASIMHAN, K. **Improving Language Understanding by Generative Pre-Training**. [S.l.: s.n.], 2018. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Acesso em: 1 fev. 2026.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. [S.l.]: arXiv, 2019. DOI: 10.48550/ARXIV.1910.10683. Disponível em: https://doi.org/10.48550/ARXIV.1910.10683. Acesso em: 1 fev. 2026.

RAITTZ, R. T.; PIERRI, C. R. D.; MALUK, M.; BATISTA, M. B.; CARMONA, M.; JUNGHARE, M.; FAORO, H.; CRUZ, L. M.; BATTISTONI, F.; SOUZA, E. de; OLIVEIRA PEDROSA, F. de; CHEN, W. M.; POOLE, P. S.; DIXON, R. A.; JAMES, E. K. Comparative genomics provides insights into the taxonomy of azoarcus and reveals separate origins of nif genes in the proposed azoarcus and aromatoleum genera. **Genes**, v. 12, p. 1–21, 2021. DOI: 10.3390/genes12010071. Disponível em: https://doi.org/10.3390/genes12010071. Acesso em: 1 fev. 2026.

RANI, J.; SHAH, A. B. R.; RAMACHANDRAN, S. **pubmed.mineR: an R package with text-mining algorithms to analyse PubMed abstracts**. en. v. 40. [S.l.]: Springer Science e Business Media LLC, out. 2015. P. 671–682. DOI: 10.1007/s12038-015-9552-2. Disponível em: https://doi.org/10.1007/s12038-015-9552-2. Acesso em: 1 fev. 2026.

ROBERTSON, S. Understanding inverse document frequency: On theoretical arguments for IDF. **Journal of Documentation**, v. 60, n. 5, p. 503–520, 2004. DOI: 10.1108/00220410410560582. Disponível em: https://doi.org/10.1108/00220410410560582. Acesso em: 1 fev. 2026.

RUBEL, E. T.; RAITTZ, R. T.; COIMBRA, N. A. d. R.; GEHLEN, M. A. C.; PEDROSA, F. d. O. ProClaT, a new bioinformatics tool for in silico protein reclassification: case study of DraB, a protein coded from the draTGB operon in Azospirillum brasilense. **BMC bioinformatics**, BMC Bioinformatics, v. 17, Suppl 18, dez. 2016. ISSN 1471-2105. DOI: 10.1186/S12859-016-1338-5. Disponível em: https://doi.org/10.1186/S12859-016-1338-5. Acesso em: 1 fev. 2026.

SAHLGREN, M. **An Introduction to Random Indexing**. [S.l.: s.n.], 2005. Disponível em: https://core.ac.uk/outputs/11433324. Acesso em: 1 fev. 2026.

SCHWARTZ, R.; DODGE, J.; SMITH, N. A.; ETZIONI, O. Green AI. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 63, n. 12, p. 54–63, nov. 2020. ISSN 0001-0782. DOI: 10.1145/3381831. Disponível em: https://doi.org/10.1145/3381831. Acesso em: 1 fev. 2026.

SCORZATO, L. Reliability and Interpretability in Science and Deep Learning. **Minds and Machines**, Springer Science e Business Media B.V., v. 34, n. 3, p. 1–31, set. 2024. ISSN 15728641. DOI: 10.1007/S11023-024-09682-0. Disponível em: https://doi.org/10.1007/S11023-024-09682-0. Acesso em: 1 fev. 2026.

SILVA FILHO, A. C. da; MARCHAUKOSKI, J. N.; RAITTZ, R. T.; PIERRI, C. R. D.; JESUS SOARES MACHADO, D. de; FADEL-PICHETH, C. M. T.; PICHETH, G. Prediction and Analysis in silico of Genomic Islands in Aeromonas hydrophila. **Frontiers in Microbiology**, v. 12, 2021. DOI: 10.3389/fmicb.2021.769380. Disponível em: https://doi.org/10.3389/fmicb.2021.769380. Acesso em: 1 fev. 2026.

SMALHEISER, N. R.; FRAGNITO, D. P.; TIRK, E. E. **Anne O'Tate: Value-added PubMed search engine for analysis and text mining**. en. v. 16. [S.l.]: Public Library of Science (PLoS), mar. 2021. e0248335. DOI: 10.1371/journal.pone.0248335. Disponível em: https://doi.org/10.1371/journal.pone.0248335. Acesso em: 1 fev. 2026.

SPITALE, G.; GERMANI, F.; BILLER-ANDORNO, N. **TopicTracker - An advanced software pipeline for text mining on PubMed data: Bridging the gap between off-the-shelf tools and code based approaches**. en. v. 10. [S.l.]: Elsevier BV, set. 2024. e36351. DOI: 10.1016/j.heliyon.2024.e36351. Disponível em: https://doi.org/10.1016/j.heliyon.2024.e36351. Acesso em: 1 fev. 2026.

TAVELLA, T. A.; CASSIANO, G. C.; COSTA, F. T. M.; SUNNERHAGEN, P.; BILSLAND, E. Yeast-based high-throughput screens for discovery of kinase inhibitors for neglected diseases. **Advances in protein chemistry and structural biology**, Adv Protein Chem Struct Biol, v. 124, p. 275–309, jan. 2021. ISSN 1876-1631. DOI: 10.1016/BS.APCSB.2020.09.007. Disponível em: https://doi.org/10.1016/BS.APCSB.2020.09.007. Acesso em: 1 fev. 2026.

TAY, Y.; DEHGHANI, M.; BAHRI, D.; METZLER, D. Efficient Transformers: A Survey. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 55, n. 6, p. 1–28, dez. 2022. ISSN 1557-7341. DOI: 10.1145/3530811. Disponível em: http://doi.org/10.1145/3530811. Acesso em: 1 fev. 2026.

THEODOSIOU, T.; DARZENTAS, N.; ANGELIS, L.; OUZOUNIS, C. A. **PuReD-MCL: a graph-based PubMed document clustering methodology**. en. v. 24. [S.l.]: Oxford University Press (OUP), set. 2008. P. 1935–1941. DOI: 10.1093/bioinformatics/btn318. Disponível em: https://doi.org/10.1093/bioinformatics/btn318. Acesso em: 1 fev. 2026.

TSHITOYAN, V.; DAGDELEN, J.; WESTON, L.; DUNN, A.; RONG, Z.; KONONOVA, O.; PERSSON, K. A.; CEDER, G.; JAIN, A. Unsupervised word embeddings capture latent knowledge from materials science literature. **Nature**, Nature Publishing Group, v. 571, n. 7763, p. 95–98, jul. 2019. ISSN 0028-0836. DOI: 10.1038/s41586-019-1335-8. Disponível em: https://doi.org/10.1038/s41586-019-1335-8. Acesso em: 1 fev. 2026.

TSURUOKA, Y.; MIWA, M.; HAMAMOTO, K.; TSUJII, J.; ANANIADOU, S. **Discovering and visualizing indirect associations between biomedical concepts**.

en. v. 27. [S.l.]: Oxford University Press (OUP), jul. 2011. P. i111–9. DOI: 10.1093/bioinformatics/btr214. Disponível em: https://doi.org/10.1093/bioinformatics/btr214. Acesso em: 1 fev. 2026.

TUDOR, C. O.; ROSS, K. E.; LI, G.; VIJAY-SHANKER, K.; WU, C. H.; ARIGHI, C. N. **Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system**. en. v. 2015. [S.l.]: Oxford University Press (OUP), mar. 2015. DOI: 10.1093/database/bav020. Disponível em: https://doi.org/10.1093/database/bav020. Acesso em: 1 fev. 2026.

TURINA, P.; FARISELLI, P.; CAPRIOTTI, E. **ThermoScan: Semi-automatic identification of protein stability data from PubMed**. en. v. 8. [S.l.]: Frontiers Media SA, mar. 2021. P. 620475. DOI: 10.3389/fmolb.2021.620475. Disponível em: https://doi.org/10.3389/fmolb.2021.620475. Acesso em: 1 fev. 2026.

VAN ROSSUM, G.; FOUNDATION, P. S. **Python Language Reference**. [S.l.: s.n.], 2026. Disponível em: https://docs.python.org/3/reference. Acesso em: 1 fev. 2026.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention Is All You Need. arXiv, 2017. DOI: 10.48550/arXiv.1706.03762. Disponível em: https://arXiv.org/abs/1706.03762. Acesso em: 1 fev. 2026.

WANG, S.; LI, B. Z.; KHABSA, M.; FANG, H.; MA, H. Linformer: Self-Attention with Linear Complexity. arXiv, 2020. DOI: 10.48550/arXiv.2006.04768. Disponível em: http://doi.org/10.48550/arXiv.2006.04768. Acesso em: 1 fev. 2026.

WEI, C.-H.; ALLOT, A.; LEAMAN, R.; LU, Z. **PubTator central: automated concept annotation for biomedical full text articles**. en. v. 47. [S.l.]: Oxford University Press (OUP), jul. 2019. w587–w593. DOI: 10.1093/nar/gkz389. Disponível em: https://doi.org/10.1093/nar/gkz389. Acesso em: 1 fev. 2026.

WILLIAMS, K.; BILSLAND, E.; SPARKES, A.; AUBREY, W.; YOUNG, M.; SOLDATOVA, L. N.; DE GRAVE, K.; RAMON, J.; DE CLARE, M.; SIRAWARAPORN, W.; OLIVER, S. G.; KING, R. D. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. **Journal of the Royal Society, Interface**, J R Soc Interface, v. 12, n. 104, mar. 2015. ISSN

1742-5662. DOI: 10.1098/RSIF.2014.1289. Disponível em: https://doi.org/10.1098/RSIF.2014.1289. Acesso em: 1 fev. 2026.

YANG, L.; ZHANG, X.; CHEN, J.; QISHAN, W.; WANG, L.; JIANG, Y.; PAN, Y. **ReCGiP, a database of reproduction candidate genes in pigs based on bibliomics**. en. v. 8. [S.l.]: Springer Science e Business Media LLC, ago. 2010. P. 96. DOI: 10.1186/1477-7827-8-96. Disponível em: https://doi.org/10.1186/1477-7827-8-96. Acesso em: 1 fev. 2026.

YAO, X.; ZHENG, Y.; YANG, X.; YANG, Z. **NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework**. [S.l.: s.n.], 2022. DOI: 10.48550/arXiv.2111.04130. Disponível em: https://doi.org/10.48550/arXiv.2111.04130. Acesso em: 1 fev. 2026.

ZHANG, X.; ZHAO, J.; LECUN, Y. **Character-level Convolutional Networks for Text Classification**. [S.l.: s.n.], 2016. DOI: 10.48550/arXiv.1509.01626. Disponível em: https://doi.org/10.48550/arXiv.1509.01626. Acesso em: 1 fev. 2026.

ZHANG, Y.; LI, M.; LONG, D.; ZHANG, X.; LIN, H.; YANG, B.; XIE, P.; YANG, A.; LIU, D.; LIN, J.; HUANG, F.; ZHOU, J. **Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models**. [S.l.: s.n.], 2025. DOI: 10.48550/arXiv.2506.05176. Disponível em: https://doi.org/10.48550/arXiv.2506.05176. Acesso em: 1 fev. 2026.