Molecular k-mer Images with Vision Transformers for Blood–Brain Barrier Permeability Prediction

Giovanna A. P. Soares[1], Hannah I. da S. Marques[1], Raquel de M. Barbosa[3]. Marcelo A. C. Fernandes[1,2,4]

[1]InovAI Lab, nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brasil

[2]Centro Multiusuário de Bioinformática (BioME) – IMD/UFRN

[3]Faculty of Pharmacy, Granada University, Granada, Spain

[4]Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, Brasil

giovanna.assuncao.705@ufrn.edu.br, hannahisabele1516@gmail.com

rbarbosa@ugr.es, mfernandes@dca.ufrn.br

Resumo. *A predição da permeabilidade da barreira hematoencefálica (BBB) é um problema importante no desenvolvimento de fármacos, pois determina se uma molécula consegue alcançar o sistema nervoso central. Este trabalho propõe uma abordagem baseada na geração de representações visuais moleculares derivadas de sequências SMILES. Inicialmente, as moléculas são tokenizadas e utilizadas para construir matrizes de contagem de k-mers (MC), posteriormente convertidas em imagens denominadas Molecule Image Representations (MIR). Essas imagens são processadas por um modelo Vision Transformer (ViT), responsável por extrair embeddings moleculares utilizados por classificadores supervisionados. Experimentos conduzidos nos datasets BBBP e B3DB demonstraram que a abordagem proposta produz resultados competitivos em termos de AUC, alcançando 0,9281 no BBBP (SVM) e 0,9125 no B3DB (RF). Os resultados indicam que os embeddings gerados pelo ViT capturam características estruturais relevantes das moléculas. Além disso, o ajuste fino parcial do modelo contribui para melhorar a qualidade das representações aprendidas. De forma geral, o método apresenta uma alternativa simples e eficaz às abordagens baseadas em grafos moleculares e descritores físico-químicos extensivos.*

Abstract. *The prediction of blood–brain barrier (BBB) permeability is an important problem in drug development, as it determines whether a molecule can reach the central nervous system. This work proposes an approach based on the generation of visual molecular representations derived from SMILES sequences. Molecules are tokenized and used to construct k-mer count matrices (MC), subsequently converted into images called Molecule Image Representations (MIR). These images are processed by a Vision Transformer (ViT) model, which extracts molecular embeddings used by supervised classifiers. Experiments on the BBBP and B3DB datasets demonstrated competitive results in terms of AUC, reaching 0.9281 on BBBP (SVM) and 0.9125 on B3DB (RF). The results indicate that ViT embeddings capture relevant structural features of molecules, and that partial*

fine-tuning improves representation quality. Overall, the method offers a simple and effective alternative to molecular graph-based and descriptor-intensive approaches.

1. Introdução

A barreira hematoencefálica (*blood–brain barrier* - BBB) é uma interface fisiológica altamente seletiva que regula a passagem de moléculas da corrente sanguínea para o sistema nervoso central. Esse mecanismo de proteção impede a entrada da maioria das pequenas moléculas no cérebro, tornando a predição de permeabilidade à BBB (*BBB permeability* - BBBp) um desafio importante no desenvolvimento de fármacos e na avaliação farmacocinética.

A capacidade de estimar se uma molécula consegue atravessar essa barreira é essencial para o desenvolvimento de terapias destinadas ao tratamento de doenças neurológicas, uma vez que compostos incapazes de atravessar a BBB não conseguem atingir seus alvos terapêuticos no sistema nervoso central. Abordagens baseadas em aprendizado de máquina (*Machine Learning* - ML) têm sido amplamente utilizadas para aprender relações entre estrutura molecular e permeabilidade a partir de dados experimentais. O *benchmark* MoleculeNet consolidou a predição de BBBp como um problema clássico de classificação em modelagem molecular, evidenciando a influência da qualidade dos dados, da diversidade química e do desbalanceamento de classes no desempenho dos modelos [Wu et al. 2018].

Desde então, diversos estudos exploraram arquiteturas baseadas em aprendizado profundo e grafos moleculares. Tong et al. introduziram estimativas de incerteza em modelos de BBBp [Tong et al. 2022], enquanto *frameworks* baseados em representações visuais, como o ImageMol, demonstraram que imagens moleculares podem servir como entrada para modelos auto-supervisionados em tarefas de ADMET [Zeng et al. 2022]. Abordagens interpretáveis baseadas em descritores de fragmentos também mostraram que representações mais simples podem capturar determinantes estruturais relevantes da permeabilidade BBB [Jia and Sosso 2024]. Nesse contexto, estratégias de tokenização quimicamente informadas (*chemistry-aware tokenization*) preservam informações estruturais importantes para a geração de *embeddings* moleculares mais expressivos [Visan and Negut 2024], e diferentes codificações moleculares — sequências, grafos ou grades bidimensionais — têm demonstrado capacidade de capturar padrões estruturais complexos [Dou et al. 2023, Guzman-Pando et al. 2024].

Outra linha de pesquisa explora representações baseadas em k-mers, nas quais padrões de coocorrência extraídos de sequências simbólicas são convertidos em estruturas bidimensionais adequadas ao processamento por redes neurais. Em genômica viral, matrizes de frequência de k-mers têm sido utilizadas para representar sequências de nucleotídeos como imagens em escala de cinza, permitindo que modelos convolucionais discriminem variantes do SARS-CoV-2 com alta precisão [Câmara et al. 2022, Coutinho et al. 2023]. Estratégias semelhantes também foram aplicadas à predição de interações fármaco–alvo, nas quais sequências moleculares ou proteicas são decompostas em k-mers sobrepostos e convertidas em matrizes bidimensionais que capturam dependências locais relevantes para modelagem de interações [De Souza et al. 2022]. Esses estudos demonstram que representações visuais derivadas de k-mers podem transformar

informações simbólicas em padrões espaciais adequados ao uso por modelos de aprendizado profundo.

Motivado por essas abordagens, este trabalho investiga um *pipeline* de representação molecular que integra tokenização simbólica de SMILES, matrizes de contagem de k-mers (MC) e representações visuais processadas por um *Vision Transformer* (ViT). Inicialmente, as sequências SMILES são convertidas em tokens químicos e utilizadas para construir matrizes de contagem de k-mers (MC), posteriormente transformadas em imagens moleculares denominadas *Molecule Image Representations* (MIR). Essas imagens são processadas por um modelo ViT previamente treinado, cujo *fine-tuning* permite extrair *embeddings* moleculares utilizados como entrada para quatro classificadores supervisionados: *Support Vector Machine* (SVM), *Extreme Gradient Boosting* (XGB), *Random Forest* (RF) e *Logistic Regression* (LR). A avaliação é realizada nos *datasets* BBBP e B3DB sob um protocolo experimental reprodutível. Os resultados indicam que SVM e XGB obtêm os maiores valores de AUC, mostrando que a combinação entre MIR e *embeddings* extraídos por ViT fornece representações discriminativas para a predição de permeabilidade BBB.

2. Metodologia

A Figura 1 apresenta uma visão geral do pipeline proposto para predição de BBB. O processo inicia com a conversão das sequências SMILES em tokens químicos, seguida da contagem de frequência de k-mers e remodelação em matrizes de contagem (MC). Essas matrizes são convertidas em imagens moleculares (MIR), que são utilizadas como entrada para um modelo ViT. Após o fine-tuning do modelo, embeddings moleculares são extraídos e utilizados como atributos de entrada para classificadores supervisionados.

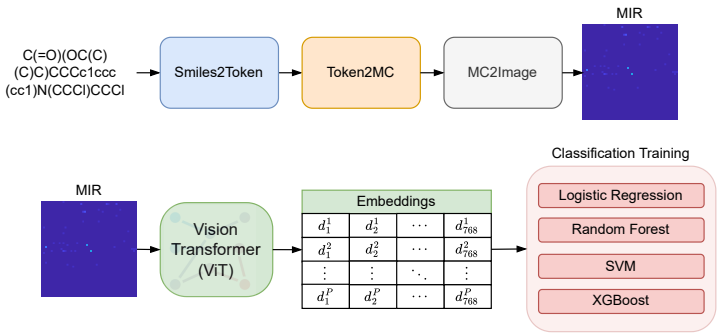


Figura 1. Pipeline completo do método proposto.

2.1. Tokenização Química de SMILES (Smiles2Token)

A primeira etapa converte representações SMILES em sequências estruturadas de tokens químicos via módulo *Smiles2Token*. Diferentemente de uma tokenização por caracteres, a abordagem *chemistry-aware* reconhece elementos de múltiplos caracteres, marcadores de aromaticidade, índices de anéis e ramificações moleculares, preservando a informação estrutural codificada no SMILES. A Tabela 1 apresenta um exemplo desse processo.

Tabela 1. Exemplo de tokenização de SMILES (módulo Smiles2Token).

Entrada SMILES
`CC1=CC=C(C=C1)C(=O)C2=CC=C(N2C)CC(=O)[O-].O.O.[Na+]`
Sequência de tokens (saída do Smiles2Token)
`{C, C, 1, =, C, C, =, C, (, C, =, C, 1, ), C, (, =, O, ), C, 2, =, C, C, =, C, (, N, 2, C, ),C, C, (, =, O, ), [O-], ., O, ., O, ., [Na+]}`

2.2. Contagem de Frequência de k-mers e Remodelação (Token2MC)

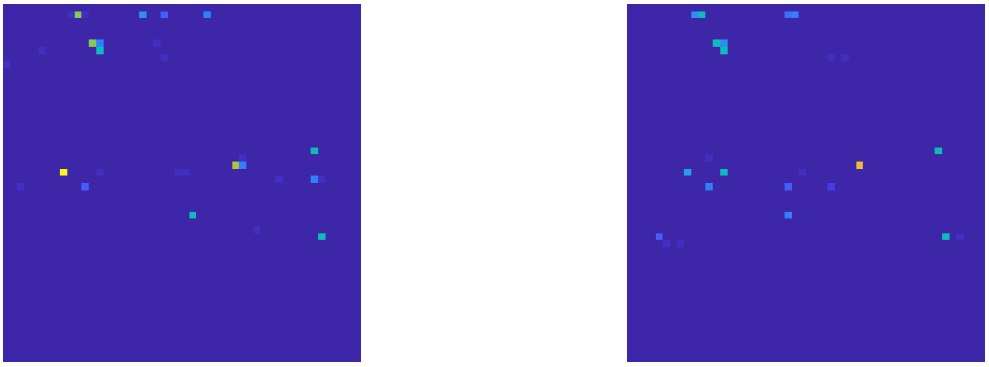
A partir da sequência de tokens químicos gerada pelo módulo *Smiles2Token*, é realizada a contagem de frequência de k-mers. Nessa etapa, os tokens são organizados em subsequências sobrepostas de comprimento k por meio de uma janela deslizante ao longo da sequência. Neste trabalho foi adotado $k = 3$, de modo que cada k-mer corresponde a três tokens consecutivos extraídos da sequência. Valores maiores de k foram avaliados preliminarmente, mas resultaram em matrizes excessivamente esparsas, nas quais a maior parte das posições permanecia nula, prejudicando a formação de padrões visuais coerentes para o processamento pelo ViT. A frequência de cada combinação possível de k-mer é contabilizada e acumulada em um vetor de comprimento fixo igual a $|V|^k$, onde $|V|$ denota o tamanho do vocabulário do tokenizador. Para os *datasets* utilizados, o tokenizador produz um vocabulário de $|V| = 71$–72 tokens, de modo que, com $k = 3$, o vetor de contagem possui comprimento $|V|^3 \approx 3{,}6 \times 10^5$. Esse vetor é então remodelado em uma matriz quadrada de contagem (MC) de dimensão $\lfloor\sqrt{|V|^k}\rfloor \times \lfloor\sqrt{|V|^k}\rfloor \approx 598 \times 598$ a 610×610 antes do redimensionamento, produzindo uma representação de dimensão fixa que depende apenas de k e do vocabulário do tokenizador, independentemente do comprimento da molécula.

2.3. Conversão Matriz–Imagem

A etapa denominada *MC2Image*, ilustrada na Figura 1, recebe como entrada a matriz de contagem (MC) e a converte em uma representação visual denominada MIR. Inicialmente, cada matriz é normalizada por escalonamento min–max, e os valores resultantes são mapeados por meio do *colormap jet*, produzindo uma imagem RGB que preserva a distribuição das frequências de k-mers em uma estrutura bidimensional colorida. O *colormap jet* foi adotado por fidelidade à estratégia *MC2Image* dos trabalhos que motivaram este estudo [Câmara et al. 2022, De Souza et al. 2022, Coutinho et al. 2023]; reconhece-se, contudo, que *jet* não é perceptualmente uniforme, e alternativas como *viridis* ou canal único replicado constituem escolhas mais defensáveis, ficando como investigação futura. Como o ViT exige entrada de dimensão fixa, a imagem é redimensionada para 224×224 pixels. A Figura 2 apresenta dois exemplos de MIRs geradas pelo processo *MC2Image*; as sequências SMILES correspondentes constam na Tabela 2.

Tabela 2. SMILES das moléculas utilizadas como exemplo na geração das MIRs.

Molécula	SMILES
Molécula 1	`C/C=C/C(C)C(O)C1C(=O)NC(CC)C(=O)N(C)CC(=O)N(C)C(CC(C)C)C(=O)NC(C(C)C)C(=O)N(C)C(CC(C)C)C(=O)NC(C)C(=O)NC(C)C(=O)N(C)C(CC(C)C)C(=O)N(C)C(CC(C)C)C(=O)N1C`
Molécula 2	`O=C(Cc1ccc(Cl)c(Cl)c1)N1CCn2ncnc2C1CN1CCCC1`



(a) Molécula 1 (b) Molécula 2

Figura 2. MIR geradas pelo processo *MC2Image*.

Os exemplos ilustram como variações estruturais nas sequências SMILES produzem padrões espaciais distintos nas MIRs: a Molécula 1, de cadeia longa, gera uma matriz de maior dimensão, enquanto a Molécula 2, menor e com anéis aromáticos, produz padrões mais compactos.

2.4. Fine-tuning do ViT

As imagens MIR geradas na etapa anterior são utilizadas como entrada para o modelo *ViT-Base-Patch16-224* [Dosovitskiy et al. 2021], pré-treinado no ImageNet-21k. A escolha do ViT decorre de duas propriedades do problema: (i) a atenção global entre *patches* captura dependências de longo alcance entre regiões distantes da matriz, relevante porque a ordenação linha/coluna dos k-mers não possui estrutura espacial local natural; e (ii) o token *[CLS]* sintetiza a imagem em um único vetor de representação global, adequado como *embedding* molecular de dimensão fixa. Os pesos pré-treinados fornecem detectores genéricos de contraste e borda que aceleram a adaptação ao domínio químico, hipótese validada pela superioridade das configurações com blocos descongelados (Seção 3). O modelo recebe imagens RGB de 224×224 pixels, divididas em *patches* de 16×16 (196 *patches*), processados por uma arquitetura *Transformer* de 12 blocos *encoder*. O token *[CLS]* produz um vetor de dimensão 768 que sintetiza as informações espaciais da MIR. Para adaptar o modelo ao domínio químico, adotou-se *fine-tuning* com congelamento seletivo, variando o número de blocos descongelados: 0 (todo o *encoder* congelado), 1 e 2. Essa estratégia investiga o impacto do grau de adaptação na qualidade das representações. O *fine-tuning* é realizado independentemente dentro de cada *fold*, garantindo ausência de vazamento de dados (*data leakage*) no protocolo descrito na Seção 2.8.

2.5. Extração de Embeddings

Após o treinamento do ViT, *embeddings* moleculares são extraídos a partir do token *[CLS]*, resultando em vetores de dimensão 768. Como o protocolo utiliza validação cruzada de cinco *folds* (Seção 2.8), cada configuração produz cinco modelos, um por *fold*. Para cada amostra do conjunto de teste — que permaneceu fora de todos os *folds* durante o treinamento — são computados cinco *embeddings*, um de cada modelo de *fold*, combinados por média elemento a elemento (*ensemble averaging*), configurando um *ensemble* padrão de k-*fold* que reduz a variabilidade estocástica. Para as amostras de treino e validação, utiliza-se o *embedding* do modelo do *fold* em que cada amostra esteve na partição de validação.

2.6. Datasets

Dois datasets públicos foram utilizados neste estudo: BBBP e B3DB. O dataset BBBP (*Blood–Brain Barrier Permeability*), disponibilizado pelo benchmark MoleculeNet, contém aproximadamente 2.050 moléculas anotadas de acordo com sua capacidade de atravessar a barreira hematoencefálica. Cada molécula é representada por uma sequência SMILES e um rótulo binário indicando se a molécula é permeável (BBB+) ou não permeável (BBB–). O dataset B3DB (*Blood–Brain Barrier Database*) [Meng et al. 2021] contém 7.807 moléculas anotadas com relação à permeabilidade BBB. Assim como no BBBP, cada amostra inclui a sequência SMILES da molécula e o rótulo binário correspondente. Ambos os datasets passam pelas mesmas etapas de pré-processamento e geração das representações MIR descritas anteriormente, garantindo consistência metodológica entre os experimentos.

2.7. Classificação Supervisionada

Os *embeddings* moleculares extraídos pelo ViT são utilizados como atributos de entrada para quatro algoritmos de aprendizado supervisionado: SVM [Cortes and Vapnik 1995], XGB [Chen and Guestrin 2016], RF [Breiman 2001] e LR. O SVM utiliza kernel RBF (*Radial Basis Function*). A escolha desses modelos cobre famílias de fronteira de decisão distintas — linear (LR), margem com kernel não linear (SVM) e árvores/ensembles (RF e XGB) — permitindo verificar se os ganhos decorrem da qualidade dos *embeddings* ou de um viés indutivo específico. Optou-se por treinar classificadores externos, em vez da cabeça de classificação do próprio ViT, para (i) isolar a avaliação da qualidade da representação e (ii) permitir comparação justa entre famílias de classificadores sobre a mesma representação fixa. Esses modelos são treinados sobre os vetores da etapa anterior, avaliando a capacidade das representações aprendidas em discriminar moléculas permeáveis (BBB+) e não permeáveis (BBB–).

Os hiperparâmetros dos classificadores foram determinados previamente via *grid search* com validação cruzada e mantidos fixos neste estudo, de modo que as variações de desempenho observadas reflitam principalmente a qualidade das representações extraídas pelo ViT. As configurações utilizadas para o dataset BBBP e B3DB são apresentadas na Tabela 3.

Tabela 3. Hiperparâmetros dos classificadores utilizados nos datasets.

Modelo	BBBP	B3DB
XGB	$\eta = 0.05$, depth=6, $n = 200$	$\eta = 0.05$, depth=6, $n = 200$
RF	$ntree = 300$, $mtry = 50$	$ntree = 500$, $mtry = 50$
LR	$\alpha = 0.4$	$\alpha = 0.1$
SVM	$cost = 1.5$, $gamma = 0.01$	$cost = 1.5$, $gamma = 0.025$

2.8. Protocolo Experimental

Os experimentos foram conduzidos seguindo um protocolo reprodutível com sementes aleatórias fixas. Foi adotada uma divisão aleatória estratificada por classe (*random split*), consistente com o protocolo utilizado pelos trabalhos com os quais este método é comparado nas Tabelas 6 e 7, o que viabiliza a comparação direta de resultados. Para cada *dataset*, as amostras foram divididas em dois conjuntos: 85% dos dados destinados ao

treinamento e validação e 15% reservados exclusivamente para teste. Sobre o conjunto de treinamento e validação foi aplicada validação cruzada estratificada de cinco *folds*, preservando a proporção entre as classes. Em cada *fold*, $1/5$ das amostras do conjunto de treinamento e validação é reservado para validação e $4/5$ para treino. O desvio padrão reportado nas tabelas de resultados reflete a variação de desempenho entre os cinco *folds*. O conjunto de teste permanece separado durante todo o treinamento e é utilizado apenas na avaliação final. O desempenho foi avaliado pela métrica AUC-ROC, amplamente utilizada em classificação binária por ser independente do limiar de decisão, complementada por curvas ROC para os conjuntos de treinamento, validação e teste.

3. Resultados

3.1. Resultados no dataset BBBP

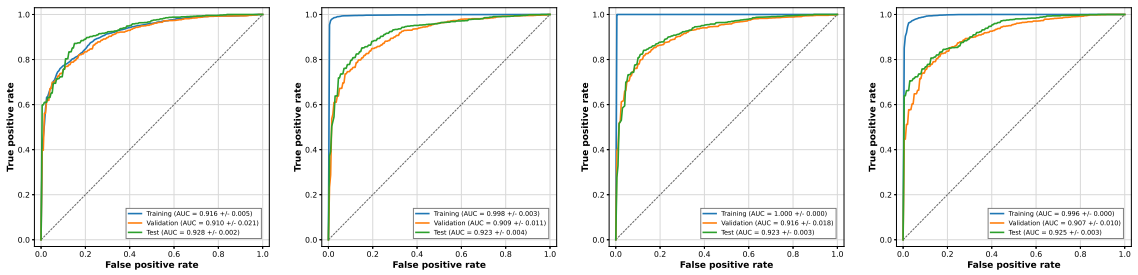
Os resultados da Tabela 4 mostram que o aumento do número de blocos descongelados no fine-tuning do ViT contribuiu, de modo geral, para a melhoria do desempenho dos classificadores no dataset BBBP. As melhores performances foram observadas para a configuração com 2 blocos descongelados, indicando que um maior grau de adaptação do modelo pré-treinado ao domínio molecular favoreceu a qualidade dos embeddings extraídos. Entre os modelos avaliados, o maior valor de AUC em teste foi obtido pelo SVM com 2 blocos descongelados (0,9281), seguido de perto pela LR (0,9249), pelo XGB (0,9228) e pelo RF (0,9226), todos com 2 blocos. Como essas diferenças são, em vários casos, menores ou da mesma ordem que os desvios padrão, não é possível afirmar que o SVM seja significativamente superior aos demais sem um teste estatístico pareado entre os *folds*; o resultado é, portanto, reportado como o de maior valor observado, e não como uma superioridade estatisticamente comprovada.

Tabela 4. Resultados para o dataset BBBP com diferentes números de blocos descongelados no fine-tuning.

Blocos	Modelo	Val AUC ± Std	Test AUC ± Std
0	SVM	0.8496 ± 0.0265	0.8946 ± 0.0080
0	XGB	0.8612 ± 0.0337	0.9030 ± 0.0074
0	RF	0.8648 ± 0.0245	0.9024 ± 0.0021
0	LR	0.8879 ± 0.0188	0.9225 ± 0.0062
1	SVM	0.8574 ± 0.0293	0.8980 ± 0.0012
1	XGB	0.8827 ± 0.0251	0.9004 ± 0.0086
1	RF	0.8815 ± 0.0205	0.9085 ± 0.0013
1	LR	0.8878 ± 0.0203	0.9247 ± 0.0042
2	SVM	0.9097 ± 0.0210	$\mathbf{0.9281 \pm 0.0019}$
2	XGB	0.9093 ± 0.0113	0.9228 ± 0.0044
2	RF	0.9161 ± 0.0180	0.9226 ± 0.0029
2	LR	0.9067 ± 0.0098	$\mathbf{0.9249 \pm 0.0030}$

Os desvios padrão foram baixos (ex.: 0,0019 para SVM com 2 blocos), indicando boa estabilidade. Observa-se ainda que as diferenças entre classificadores diminuem com o aumento dos blocos descongelados: com 2 blocos, todos atingem AUCs próximas, reforçando que a qualidade da representação aprendida pelo ViT é o fator central no desempenho final.

A Figura 3 mostra que todos os classificadores apresentam curvas ROC bem acima da diagonal, com comportamento consistente entre treinamento, validação e teste — evidenciando boa capacidade de generalização dos *embeddings* gerados pelo ViT.



(a) SVM **(b) XGB** **(c) RF** **(d) LR**

Figura 3. Curvas ROC dos classificadores para o dataset BBBP.

3.2. Resultados no dataset B3DB

Os resultados apresentados na Tabela 5 indicam que o aumento do número de blocos descongelados no fine-tuning do ViT tende a melhorar o desempenho dos classificadores no dataset B3DB. Em particular, a configuração com dois blocos descongelados apresentou os melhores resultados gerais, sugerindo que um maior grau de adaptação do modelo pré-treinado ao domínio molecular contribui para a geração de representações mais discriminativas. Entre os modelos avaliados, o melhor desempenho em teste foi obtido pelo RF com dois blocos descongelados, alcançando AUC de 0,9125. Esse resultado também corresponde ao maior valor observado no conjunto de validação (0,9250), indicando consistência entre as etapas de validação e teste. Os modelos XGB e RF apresentaram desempenho competitivo nas diferentes configurações avaliadas, mantendo valores de AUC próximos ou superiores a 0,90 nas melhores configurações.

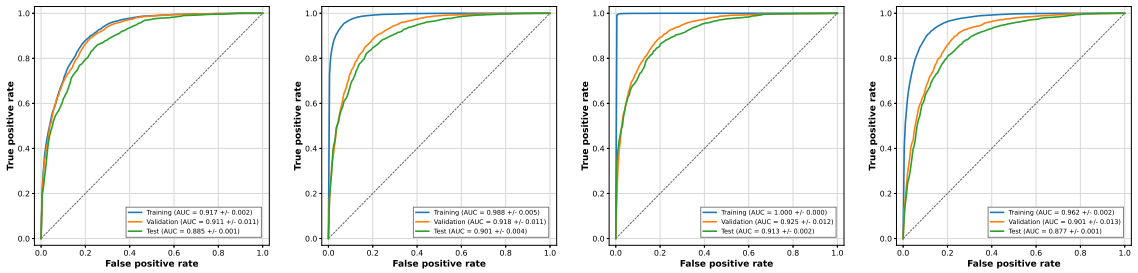
Tabela 5. Resultados para dataset B3DB com diferentes classificadores e números de blocos descongelados no fine-tuning do ViT.

Blocos	Modelo	Val AUC ± Std	Test AUC ± Std
0	SVM	0.7718 ± 0.0183	0.7750 ± 0.0023
0	XGB	0.8826 ± 0.0110	0.9002 ± 0.0022
0	RF	0.8884 ± 0.0126	0.8954 ± 0.0017
0	LR	0.8708 ± 0.0086	0.8824 ± 0.0029
1	SVM	0.8577 ± 0.0146	0.8541 ± 0.0010
1	XGB	0.8921 ± 0.0199	0.8947 ± 0.0096
1	RF	0.8997 ± 0.0155	0.9031 ± 0.0017
1	LR	0.8783 ± 0.0098	0.8825 ± 0.0028
2	SVM	0.9112 ± 0.0110	0.8851 ± 0.0009
2	XGB	0.9181 ± 0.0110	0.9007 ± 0.0040
2	RF	**0.9250 ± 0.0123**	**0.9125 ± 0.0020**
2	LR	0.9009 ± 0.0135	0.8770 ± 0.0013

Por outro lado, o SVM apresentou desempenho inferior quando nenhum bloco do ViT foi descongelado, indicando maior dependência da qualidade das representações geradas. Entretanto, ao aumentar o número de blocos ajustados durante o fine-tuning, observa-se melhora significativa em seu desempenho. Já a Logistic Regression apresentou resultados mais modestos no dataset B3DB, sugerindo que modelos lineares podem explorar de forma menos eficaz as estruturas não lineares presentes nos embeddings gerados pelo Vision Transformer. De forma geral, os resultados indicam que a qualidade das representações moleculares aprendidas pelo Vision Transformer desempenha papel fundamental no desempenho final dos classificadores, com melhorias consistentes à medida

que o modelo é progressivamente adaptado ao domínio das moléculas.

A Figura 4 confirma o bom desempenho dos classificadores no B3DB. Os modelos baseados em árvores (RF e XGB) apresentam maior capacidade de discriminação, explorando de forma mais eficaz as estruturas não lineares dos *embeddings* nesse conjunto maior.



(a) SVM **(b) XGB** **(c) RF** **(d) LR**

Figura 4. Curvas ROC dos classificadores para o dataset B3DB.

3.3. Discussão e Comparação com o Estado da Arte

A diferença de desempenho entre os classificadores nos dois *datasets* pode ser atribuída ao volume de dados: no BBBP (menor), o SVM com kernel RBF explorou eficazmente a estrutura não linear dos *embeddings*; no B3DB (cerca de quatro vezes maior), o RF destacou-se por capturar interações complexas entre as dimensões da representação.

Para o *dataset* BBBP (ver Tabela 6), o método proposto alcançou uma AUC de $0{,}9281 \pm 0{,}0019$ utilizando um protocolo de divisão aleatória estratificada dos dados (*random split*). Sob o mesmo protocolo de divisão aleatória, esse resultado é próximo, porém inferior, ao de MoleculeFormer (AUC = 0,9320), uma arquitetura *transformer* especializada em modelagem molecular, e supera D-MPNN (AUC = 0,9100). Métodos avaliados sob *scaffold split*, como AttentiveFP (AUC = 0,9200), não são diretamente comparáveis: o *scaffold split* é um protocolo mais rigoroso, que impede a sobreposição de esqueletos moleculares entre treino e teste, de modo que valores sob *random split* tendem a ser otimistas. Assim, a alegação de competitividade restringe-se a métodos avaliados sob *random split*, e não deve ser estendida aos demais. A comparação com métodos sob protocolos distintos é apresentada na tabela apenas para contextualizar a literatura, com o protocolo de cada trabalho explicitado.

Tabela 6. Comparação com outros métodos para o dataset BBBP, agrupados por protocolo de divisão dos dados.

Método	Tipo	Split	AUC
Random split			
D-MPNN [Yang et al. 2019]	GNN	Random	0.9100
MoleculeFormer [Qin et al. 2025]	Transformer	Random	0.9320
Método proposto	ViT + SVM	Random	**0.9281 ± 0.0019**
Scaffold split			
RF [Wu et al. 2018]	ML clássico	Scaffold	0.7120
GraphConv [Wu et al. 2018]	GNN	Scaffold	0.6900
AttentiveFP [Xiong et al. 2019]	GNN com atenção	Scaffold	0.9200
Outros protocolos			
MolBERT [Fabian et al. 2020]	Transformer	ChemBench com folds predefinidos	0.7612

No caso do *dataset* B3DB (ver Tabela 7), o método proposto alcançou uma AUC de 0,9125 $\pm$ 0,0020. Esse valor é inferior aos melhores resultados reportados na literatura recente, obtidos por abordagens baseadas em *ensembles* e descritores moleculares (AUC até 0,96), e também inferior ao graphB3 (AUC = 0,943 $\pm$ 0,004). A contribuição do método no B3DB não é, portanto, superar o estado da arte em valor absoluto, mas demonstrar que uma representação visual simples derivada de k-mers, sem descritores físico-químicos explícitos nem arquitetura especializada em grafos, atinge desempenho na faixa de modelos de aprendizado profundo como MegaMolBART (AUC = 0,8800). Cabe destacar que os métodos que reportam os maiores valores de AUC para o B3DB utilizam grandes conjuntos de descritores físico-químicos previamente calculados ou *ensembles* de modelos tradicionais, o que pode introduzir maior complexidade na engenharia de atributos e no treinamento dos modelos. Em contraste, a abordagem proposta utiliza diretamente representações moleculares derivadas de sequências e imagens, reduzindo a dependência de descritores químicos explícitos.

Tabela 7. Comparação com outros métodos para o dataset B3DB, agrupados por protocolo de divisão dos dados.

Método	Tipo	Split	AUC
Random split			
MegaMolBART [Huang et al. 2024]	Transformer	Random	0.8800
[Jafarpour et al. 2024]	Extra Trees + Mordred	Random	0.9500
[Jafarpour et al. 2024]	Ensemble	Random	0.9600
Método proposto	ViT + RF	Random	**0.9125 $\pm$ 0.0020**
Outros protocolos			
graphB3 [Kumar et al. 2025]	GNN	Stratified + multiple seeds	0.943 $\pm$ 0.004

De forma geral, sob o protocolo de divisão aleatória, o método proposto apresenta desempenho comparável aos métodos sob o mesmo protocolo no BBBP e permanece abaixo dos melhores resultados no B3DB, sem superar o estado da arte. O argumento central deste trabalho não é, portanto, a superioridade absoluta, mas a simplicidade do *pipeline*: uma representação visual derivada diretamente de sequências SMILES, sem descritores físico-químicos explícitos nem engenharia de grafos moleculares, é capaz de alcançar desempenho na faixa de arquiteturas de aprendizado profundo especializadas. As comparações com métodos que utilizam *scaffold split* (RF, GraphConv e AttentiveFP na Tabela 6) e outros protocolos distintos devem ser interpretadas com cautela e apenas como contextualização da literatura, uma vez que o *scaffold split* é um protocolo mais rigoroso para avaliar generalização a moléculas estruturalmente distintas das utilizadas no treinamento.

4. Conclusão

Este trabalho apresentou uma abordagem para predição da permeabilidade da barreira hematoencefálica baseada na geração de representações visuais moleculares derivadas de sequências SMILES e processadas por ViT. O método proposto transforma contagens de frequência de k-mers em imagens moleculares (MIR), permitindo que modelos de visão computacional capturem características estruturais relevantes das moléculas. Nos datasets BBBP e B3DB, sob divisão aleatória estratificada, a abordagem alcançou AUC de 0,9281 (SVM) e 0,9125 (RF), respectivamente — desempenho comparável a métodos sob o mesmo protocolo no BBBP, porém abaixo do estado da arte no B3DB. O ajuste fino parcial do ViT contribuiu consistentemente para melhorar a qualidade das representações

aprendidas, e a extração de *embeddings* seguida de classificadores externos mostrou-se eficaz, evidenciando que a informação discriminativa reside na representação visual de k-mers.

Como trabalhos futuros, planeja-se: (i) reproduzir os experimentos sob *scaffold split*; (ii) incluir *baselines* baseados em vetores de k-mers diretos e *fingerprints* clássicas (ECFP4/MACCS); (iii) comparar o ViT pré-treinado com um ViT inicializado aleatoriamente para quantificar a contribuição da transferência do ImageNet; e (iv) complementar a AUC-ROC com F1 e MCC diante do desbalanceamento das classes.

Agradecimentos

Este trabalho foi financiado pelo CNPq, pelo Sistema Único de Saúde (SUS) e pelo Ministério da Saúde (MS), por meio do projeto nº 444306/2023-4 — *PharmaGenoNet*: Uma Plataforma Integrada de Farmacogenômica Populacional e *Deep Learning* para Predição de Interações Fármaco-Alvo.

Referências

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Coutinho, M. G. F., Câmara, G. B. M., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). Sars-cov-2 virus classification based on stacked sparse autoencoder. *Computational and Structural Biotechnology Journal*, 21:284–298.

Câmara, G. B. M., Coutinho, M. G. F., Silva, L. M. D. d., Gadelha, W. V. d. N., Torquato, M. F., Barbosa, R. d. M., and Fernandes, M. A. C. (2022). Convolutional neural network applied to sars-cov-2 sequence classification. *Sensors*, 22(15):5730.

De Souza, J. G., Fernandes, M. A., and de Melo Barbosa, R. (2022). A novel deep neural network technique for drug–target interaction. *Pharmaceutics*, 14(3):625.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Dou, B., Zhu, Z., Merkurjev, E., Ke, L., Chen, L., Jiang, J., Zhu, Y., Liu, J., Zhang, B., and Wei, G.-W. (2023). Machine learning methods for small data challenges in molecular science. *Chemical Reviews*, 123(13):8736–8780.

Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., and Ahmed, M. (2020). Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Guzman-Pando, A., Ramirez-Alonso, G., Arzate-Quintana, C., and Camarillo-Cisneros, J. (2024). Deep learning algorithms applied to computational chemistry. *Molecular Diversity*, 28(4):2375–2410.

Huang, E. T., Yang, J.-S., Liao, K. Y., Tseng, W. C., Lee, C., Gill, M., Compas, C., See, S., and Tsai, F.-J. (2024). Predicting blood–brain barrier permeability of molecules with a large language model and machine learning. *Scientific reports*, 14(1):15844.

Jafarpour, S., Asefzadeh, M., and Aboutaleb, E. (2024). The application of machine learning in predicting the permeability of drugs across the blood brain barrier. *Iranian Journal of Pharmaceutical Research: IJPR*, 23(1):e149367.

Jia, H. and Sosso, G. C. (2024). Transparent machine learning model to understand drug permeability through the blood–brain barrier. *Journal of Chemical Information and Modeling*, 64(23):8718–8728.

Kumar, S., Yadav, S., Vora, D. S., Paul, P., and Dhanjal, J. K. (2025). graphb3—an interpretable graph learning approach for predicting blood–brain barrier permeability. *Briefings in Bioinformatics*, 26(6):bbaf679.

Meng, F., Xi, Y., Huang, J., and Ayers, P. W. (2021). B3DB: A multiscale dataset for understanding and predicting the blood–brain barrier permeability. *Scientific Data*, 8(1):289.

Qin, M., Sun, Z., Feng, L., Han, C., Xia, J., and Han, L. (2025). Moleculeformer is a gcn-transformer architecture for molecular property prediction. *Communications Biology*, 8(1):1668.

Tong, X., Wang, D., Ding, X., Tan, X., Ren, Q., Chen, G., Rong, Y., Xu, T., Huang, J., Jiang, H., et al. (2022). Blood–brain barrier penetration prediction enhanced by uncertainty estimation. *Journal of Cheminformatics*, 14(1):44.

Visan, A. I. and Negut, I. (2024). Integrating artificial intelligence for drug discovery in the context of revolutionizing drug delivery. *Life*, 14(2):233.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388.

Zeng, X., Xiang, H., Yu, L., Wang, J., Li, K., Nussinov, R., and Cheng, F. (2022). Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016.