Definições

DeepMol: O DeepMol é uma ferramenta, modular e open source, de AutoML desenvolvida especificamente para o domínio da química computacional. O seu objetivo é simplificar e otimizar o processo de construção de modelos de ML e DL, com suporte a modelos convencionais como redes neurais convulsionais, recorrentes e de grafos, para prever propriedades e atividades de moléculas.   
A ferramenta foi testada em 22 conjuntos de dados de benchmark do Therapeutics Data Commons (TDC) e obteve resultados competitivos, tendo em particular destaque em tarefas de absorção e metabolismo.

BERT (Bidirectional Encoder Representations from Transformers):

É um modelo de processamento de linguagem natural (NLP) desenvolvido pela Google.

-Como é que funciona:  
O BERT é alimentado por uma poderosa arquitetura de rede-neural conhecida como Transformers. Essa arquitetura incorpora um mecanismo chamado “self-attention”, que permite ao BERT medir a importância de cada palavra com base no seu contexto, tanto anterior como posterior, que é a sua principal inovação, o treino bidirecional. Esta noção/consciência de contexto confere a capacidade de gerar embeddings contextualizados de palavras, que são a representação das palavras considerando o seu significado dentro da frase.

-Como é que é treinado:  
Treino do BERT:   
Masked Language Model (MLM): Durante o treino algumas palavras/caracteres da frase/string são escondidas (substituídas por um token [MASK]) e o modelo aprende a prever essas palavra/caracteres com base no contexto ao redor. Isto permite ao BERT ser bidirecional.

>Frase original: “O cão está a dormir na cama”  
>Entrada para o BERT: “O cão está a [MASK] na cama”  
>Output esperado: “dormir”

~~2. Next Sentence Prediction (NSP): Tem o objetivo de permitir que o BERT aprenda para além das relações entre palavras dentro de uma frase, mas também as relações entre as frases dentro de um texto. Durante o treino o BERT recebe duas frases, provenientes de um corpo de texto, que vão ser separadas por um token.   
O BERT aprende a prever se a segunda frase do par segue a logica da primeira ou não.~~

-BERT Embeddings:  
O interesse no BERT deve-se à sua capacidade de representar palavras com alta especificidade e contexto.

>Word Embeddings vs Contextual Word Embeddings: O word embeddings é a atribuição de um código que vai representar palavras. O BERT conseguiu dar um step up com o word contextual embedding em vez de atribuir um único código fixo para cada palavra vai criar diferentes representações para a mesma palavra. Isso torna a representação das palavras mais rica e influenciada pelas palavras que a rodeiam.  
  
>WordPiece Tokenization: O vocabulário do BERT funciona como um puzzle composto por subwords. Ele usa a wordpiece tokenization para dividir as palavras em subcomponentes. Isto é especialmente útil para lidar com palavras complexas e com palavras novas.

-Variantes do BERT

SMILES (Simplified Molecular Input Line Entry System):

É uma forma compacta de representar estruturas químicas de moléculas através de texto. Permite o fácil armazenamento e manipulação dos compostos por computadores.

-Como é que funciona:  
O SMILES representa átomos, ligações e estruturas químicas usando notação linear.  
>Átomos são representados pelos seus símbolos químicos:  
-Carbono (C), Oxigenio (O), Hidrogénio (H)

>Ligações químicas podem ser representadas por diferentes símbolos:  
-Simples: -  
-Dupla: =  
-Tripla: #

>Ramificações são indicadas por parenteses:  
CC(0)C

>Ciclos são numerados e conectados pelo mesmo número  
C1CCCC1

>Estereoisometria pode ser representada com símbolos como / e \ para configurações de dupla ligação ou @ para quiralidade.

Global Similarity:

É a avaliação da similaridade entre dois objetos, conjuntos de dados, imagens ou outras entidades, considerando-as como um todo em vez de especificar características individuais. Esta abordagem procura encontrar a semelhança geral entre os elementos comparados tendo em conta a estrutura, contexto ou distribuição das características.

Local Similarity:

Comparação ou avaliação da semelhança entre partes especificas de dois objetos, conjuntos de dados, imagens ou outras entidades, em vez de considerar o todo. Esta abordagem é útil quando existem regiões ou características especificas relevantes para a analise

Global vs local

>Global: Avalia o todo, mostrando a semelhança geral   
>Local: Foca em zonas especificas, identificando semelhanças pontuais.

Similaridade Biossintetica:

É uma abordagem que vai além da similaridade estrutural e química, baseando-se na comparação das moléculas com base em como estas são sintetizadas nos sistemas biológicos.   
A similaridade biossintética é calculada tendo em consideração as vias metabólicas e as enzimas envolvidas na produção de compostos naturais. Esta abordagem permite identificar relações funcionais entre moléculas que podem não ser evidentes apenas com uma análise estrutural. Falar na molecular fingerprint ?

Machine Learning: É um ramo da Inteligência artificial, com algumas variantes e aplicações, que se concentra no desenvolvimento de modelos que permite aos computadores aprenderem com dados e fazerem previsões ou tomarem decisões sem serem explicitamente programados para isso. Em vez de seguirem instruções passo a passo os sistemas de machine learning aprendem com dados ao identificar padrões e a generalizar para novos dados.

Neural Network:   
É um modelo computacional inspirado no funcionamento do cérebro humano, composto por unidades interconectadas chamadas de neurónios artificiais. Estes neurónios estão organizados em camadas e trabalham em conjunto para aprender padrões em dados. Neural networks são a base do Deep Learning.

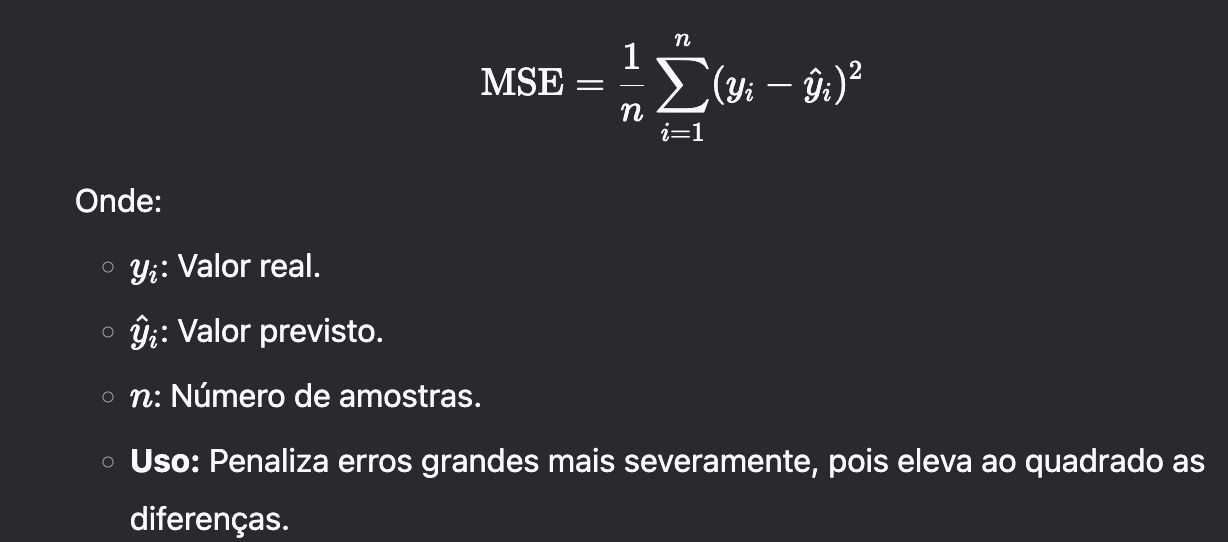
Estrutura básica de uma NN:  
- Input layer:  
Recebe os dados brutos   
Cada neurónio nesta camada representa uma característica dos dados

-Hidden layers:  
Realizam a maior parte do processamento.  
Cada neurónio numa hidden layer recebe o input dos neurónios da camada anterior, aplica uma transformação matemática e passa a informação para a próxima camada.

-Output layer:  
Produz o resultado do modelo.

Funcionamento de um neurónio artificial: (Completar)  
  
>Redes Neurais Tradicionais:  
Podem ter apenas uma ou poucas camadas ocultas.  
São eficazes para problemas simples.

>Redes Neuras Profundas:  
Possuem muitas camadas ocultas.  
Capaz de aprender representações abstratas de dados  
Requer grandes quantidades de dados e poder computacional.

Loss Function: É uma métrica usada para quantificar a performance do modelo em relação aos dados reais. Ela mede a diferença entre previsões feitas pela rede neural e os valores reais dos dados. O principal objetivo da loss function é guiar o processo de treino da rede neural, fornecendo um sinal que indique o qual errado o modelo esta para poder ser otimizado.  
Existem várias loss functions, como por exemplo:  
>MSE - Mean Squared Error  
>MAE – Mean Absolut Error  
  
Uma imagem com texto, Tipo de letra, captura de ecrã, preto

Descrição gerada automaticamente

>Binary Cross-Entropy  
Uma imagem com texto, Tipo de letra, captura de ecrã, preto

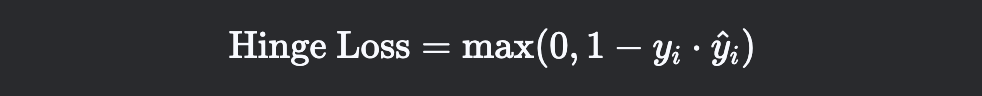
Descrição gerada automaticamente

>Categorical Cross-Entropy:

Uma imagem com texto, Tipo de letra, preto, captura de ecrã

Descrição gerada automaticamente

>Hinge Loss:



A escolha da loss function certa depende do tipo de problema e das características dos dados. Uma boa loss function deve ser diferenciável para permitir o uso de algoritmos descendentes, deve refletir adequadamente o desempenho do modelo e deve ser fácil de interpretar.

Deep Learning:  
É um subcampo do machine learning que utiliza redes neurais profundas, dai o termo deep, para aprender representações complexas e hierárquicas dos dados.

Referencias:

1. **Correia, J., Capela, J., & Rocha, M. (2024). Deepmol: an automated machine and deep learning framework for computational chemistry. *Journal of Cheminformatics, 16*(136).**[**https://doi.org/10.1186/s13321-024-00937-7**](https://doi.org/10.1186/s13321-024-00937-7)
2. **Bolcato, G., Heid, E., & Boström, J. (2022). On the value of using 3D shape and electrostatic similarities in deep generative methods. *Journal of Chemical Information and Modeling, 62*(6), 1388–1398.**[**https://doi.org/10.1021/acs.jcim.1c01535**](https://doi.org/10.1021/acs.jcim.1c01535)
3. **Zheng, X., & Tomiura, Y. (2024). A BERT-based pretraining model for extracting molecular structural information from a SMILES sequence. *Journal of Cheminformatics, 16*(71).**[**https://doi.org/10.1186/s13321-024-00848-7**](https://doi.org/10.1186/s13321-024-00848-7)
4. **Nollen, L.-M., Meijer, D., Sorokina, M., & van der Hooft, J. J. J. (2024). Biosynfoni: A Biosynthesis-informed and Interpretable Lightweight Molecular Fingerprint. *ChemRxiv*.**[**https://doi.org/10.26434/chemrxiv-2025-cwg74**](https://doi.org/10.26434/chemrxiv-2025-cwg74)
5. **Menke, J., Massa, J., & Koch, O. (2021). Natural product scores and fingerprints extracted from artificial neural networks. *Computational and Structural Biotechnology Journal, 19*, 4593–4602.**[**https://doi.org/10.1016/j.csbj.2021.07.032**](https://doi.org/10.1016/j.csbj.2021.07.032)