Dia 2

Módulo 2: Representações Numéricas e Fingerprints

Por que não usamos uma "foto" da molécula?

Modelos de **Regressão/Classificação**: Exigem entradas de tamanho fixo (vetores 1 X N)

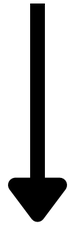
Limitações da Imagem:

- Invariância rotacional (girar a molécula não muda a química, mas muda os pixels).
- Esparsidade de dados.
- Perda de informação 3D/eletrônica explícita.

Limitações do Grafo (SMILES): O computador não entende nativamente o que é "C" ou "=".

Solução: Precisamos de um tradutor que converta a estrutura química em uma lista de números.

O "Código de Barras" da Molécula



Definição: Vetores binários (0 ou 1) que indicam a presença ou ausência de subestruturas específicas.

Bit 1 (Ligado): A subestrutura existe na molécula.

Bit 0 (Desligado): A subestrutura não existe.

Tipos Comuns:

- **Baseados em chaves** (MACCS Keys): Procura grupos funcionais pré-definidos (tem anel? tem oxigênio?).
- **Baseados em caminhos/circular** (Morgan/ECFP): Gera fragmentos baseados na conectividade.

Morgan Fingerprints (Circular Fingerprints)

Padrão ouro na indústria farmacêutica (também chamados de ECFP - Extended-Connectivity Fingerprints).

Como funciona:

1. Olha para cada átomo central.
2. Analisa os vizinhos em um Raio (r).
3. Aplica uma função de Hashing para gerar um número único para aquele ambiente químico.
4. Mapeia esse número em um vetor de tamanho fixo (ex: 1024 ou 2048 bits).

ECFP4: Significa um fingerprint circular com diâmetro 4 (R = 2). É o que usaremos na prática.

Vetores de Bits vs. Contagem

Bit Vector (O mais comum):

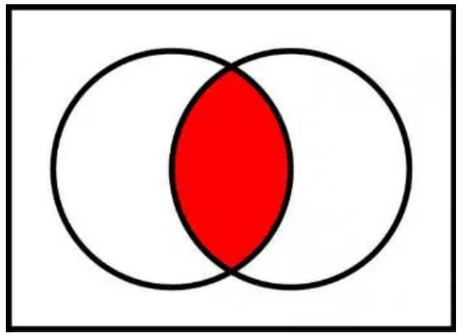
- Responde: "Este fragmento existe?" (Sim/Não).
- Vantagem: Rápido, eficiente para similaridade.
- RDKit: GetMorganFingerprintAsBitVect.

Count Vector:

- Responde: "Quantas vezes este fragmento aparece?".
- Vantagem: Útil para polímeros ou quando a repetição de grupos afeta a propriedade (ex: toxicidade).
- RDKit: GetMorganFingerprint.

Medindo Similaridade: Coeficiente de Tanimoto

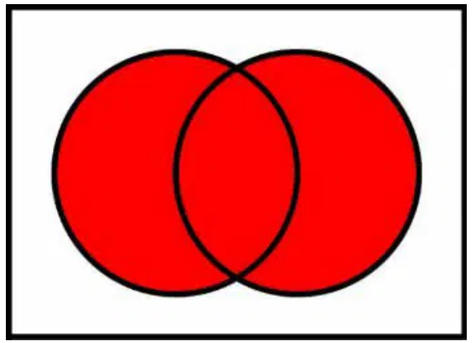
Como saber se duas moléculas são parecidas matematicamente?



Fórmula (Interseção sobre União):

$$T(A, B) = Nc / (Na + Nb - Nc)$$

- Nc: Bits ligados em comum (AND).
- Na, Nb: Total de bits ligados em A e B.

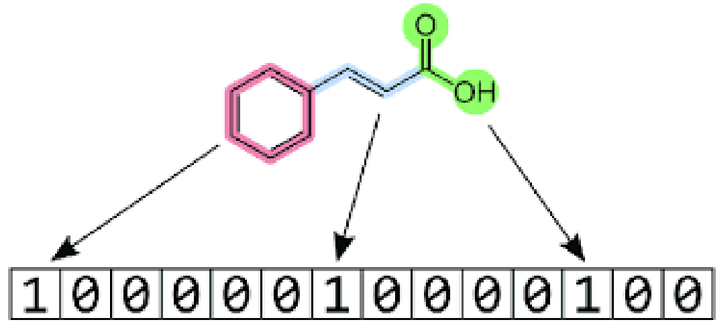


Interpretação: 1.0 = Idêntico; 0.0 = Totalmente diferente.

Cutoff: Geralmente $T > 0.7$ ou 0.8 indica alta similaridade estrutural.

Gerando Fingerprints com RDKit

1. **Gerador**: Instanciar rdFingerprintGenerator (Raio 2, 2048 bits).
2. **Cálculo:** Aplicar .GetFingerprint(mol) em cada molécula do dataset.
3. **Conversão**: O RDKit gera um objeto especial. Precisamos **converter para numpy.array** para usar no Scikit-Learn/TensorFlow.



Antes do Modelo: Entendendo os Dados

- Já calculamos **descritores** (MW, LogP e outros descritores) no Módulo 1.
- **Multicolinearidade**: Variáveis muito correlacionadas atrapalham modelos lineares.
- **Heatmap** (Mapa de Calor): Visualiza a matriz de correlação.
- Exemplo: É provável que "**Número de Átomos**" e "**Peso Molecular**" tenham correlação alta(são redundantes).