Dia 2

Módulo 1: Representação e Manipulação de Dados Moleculares

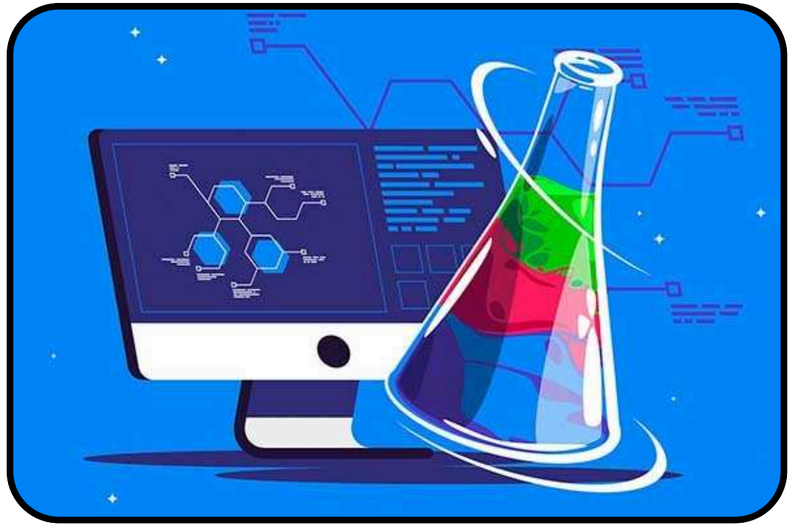
Espaço Químico e o Gargalo da Descoberta

Espaço Químico estimado de 10^{60} Moléculas

Moléculas sintetizadas na história $\approx 10^{8}$

O problema: A triagem experimental e a simulação quântica são computacionalmente custosas para varrer esse espaço.

Solução via IA: Modelos Substitutos (Surrogate Models) que aprendem padrões a partir de dados existentes para prever propriedades e frações de segundos.



Como ensinar Química para o Computador?

Computadores processam vetores e matrizes, não “desenhos”

Níveis de Representação:

- 1D (Linear): SMILES, InChI, SELFIES
- 2D (Grafo): Nós (Átomos) e Arestas (Ligações). Topologia.
- 3D (Geometria): Coordenadas Cartesianas (x,y,z) e conformeros.

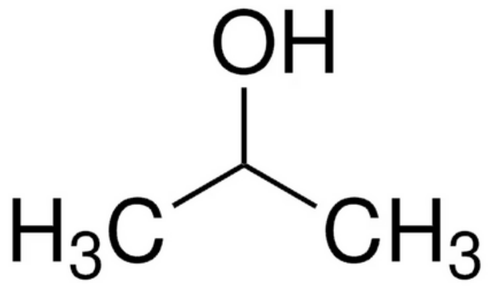
SMILES (Simplified Molecular Input Line Entry System)

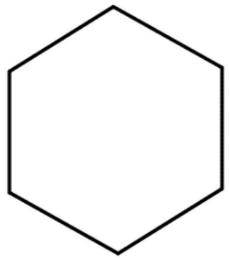
Notação linear baseada em caracteres ASCII

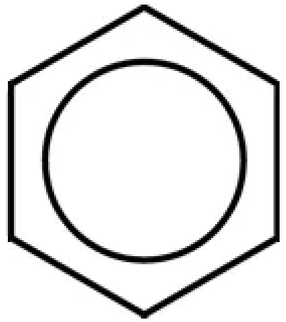
Regras Básicas:

- Átomos: C,N,O (H é implícito na valência).
- Ramificações: Parênteses ().
- Anéis: Números de fechamento.
- Aromaticidade: Letras minúsculas (c,n).

SMILES (Simplified Molecular Input Line Entry System)







Isopropanol CC(0)C Ciclohexano C1CCCCC1 Benzeno c1ccccc1

SMILES Canônico x Genérico

Uma molécula, **múltiplos** SMILES possíveis (dependendo do átomo de início).

- Exemplos Etanol: CCO, OCC, C(O)C

Algoritmos de Canonização: gera uma **string única** e determinística para a mesma estrutura topológica.

Importância: Fundamental para **remover duplicatas** em Datasets de Machine Learning

SMILES Isomérico

A química não é plana: a importância da **quiralidade** (ex: Talidomida).

Notação Isomérica:

- @ / @@: **Quiralidade tetraédrica** (Sentido horário/anti-horário).
- / / \: **Isomeria Geométrica** (Cis/Trans em duplas ligações).

Impacto nos modelos preditivos de atividade biológica.

RDKit: A "Faca Suíça" da Quimioinformática

Biblioteca Open-Source

O **Objeto Mol**: Não é uma imagem. É um objeto computacional contendo a tabela de conectividade (Grafo Molecular).

Funcionalidades:

- **Leitura/Escrita** (SMILES → Mol → PDB/SDF).
- **Cálculo de Propriedades** (Descritores).
- **Limpeza e Sanitização de estruturas** (valência incorreta, cargas).

De Estruturas para Números

Para aplicar ML, precisamos de **features** (variáveis).

Descritores 0D/1D: Contagem de átomos, Peso Molecular, Carga total.

Descritores 2D (Topológicos): LogP, TPSA (Área de Superfície Polar Topológica), Regra de Lipinski.

Descritores 3D: Volume, Momento de Dipolo (requer conformeros).

No Módulo 1, focaremos nos descritores físico-químicos clássicos.

Avaliando a "Oralidade" (Drug-Likeness)

Diretriz empírica para avaliar se um composto químico tem propriedades para ser um fármaco oral.

A **Regra dos 5** (Ro5):

- MW $\leqslant$ 500 Da.
- LogP $\leqslant$ 5 (Lipofilicidade).
- Doadores de H (HBD) $\leqslant$ 5.
- Aceitadores de H (HBA) $\leqslant$ 10.

Roteiro da Prática: Do PubChem ao Pandas

1. Coleta e Representação:
   - Busca automática via PubChemPy.
   - Conversão: String SMILES → Objeto Mol (RDKit).

2. Análise de Propriedades (Lipinski)
   - Cálculo de descritores: Peso Molecular, LogP, Doadores/Aceitadores de H.
   - Filtro de "Drug-likeness" (Regra dos 5) via código.

3. Química Virtual (SMARTS)
   - Definição de padrões de reação com linguagem SMARTS.
   - Exemplo: Hidrogenação (Etileno → Etano).
   - Desafio: Redução da Acetona → Isopropanol.

4. Organização de Dados
   - Estruturação do Dataset Final.
   - Consolidação de moléculas e propriedades em DataFrame para alimentar modelos de IA.