Dia 2

Módulo 3: Pré-processamento e Construção de Modelos

O que é QSAR?

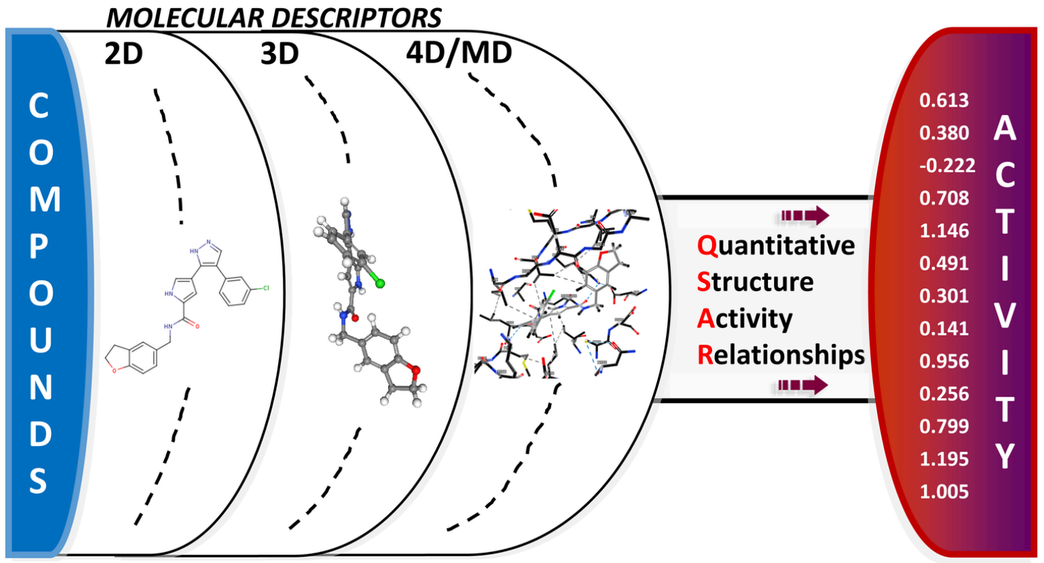
Quantitative Structure-Activity Relationship.

A **Hipótese Fundamental**: "Estruturas químicas semelhantes tendem a ter propriedades biológicas/físico-químicas semelhantes."

O Objetivo: Criar uma função matemática f(x) onde:

- **x** = Características da Molécula (Fingerprints, Descritores).
- **y** = Propriedade (Atividade, Toxicidade, Solubilidade).

Por que usar IA? As relações entre estrutura e atividade são frequentemente complexas demais para equações simples feitas à mão.



A Realidade dos Bancos de Dados

Bancos públicos (ChEMBL, PubChem) agregam dados de milhares de fontes.

Ruído Químico:

- Misturas: Solventes cristalizados junto com o fármaco.
-
- Sais: Muitos fármacos são comercializados como sais (ex: Cloridrato, Sulfato) para melhorar a solubilidade.

Impacto: Se treinarmos o modelo com "Cloridrato de X", o modelo pode aprender que o íon Cloro é importante, gerando **falso-positivos.**

Preparando a Entrada (Features X)

Modelos de IA **não entendem átomos**, apenas números.

Input (X): Matriz de Fingerprints (Módulo 2).

- Cada linha = Uma molécula.
- Cada coluna = Um bit (presença/ausência de subestrutura).

Target (y): A propriedade que queremos prever (ex: valor de pIC50).

.

Treino vs. Teste: Evitando a "Decoreba"

O Perigo do **Overfitting**: O modelo "**decora**" os exemplos conhecidos mas falha em novos compostos.

A Solução (**Split**):

- **Treino** (80%): Dados usados para o modelo aprender os padrões.
- **Teste** (20%): Dados "escondidos" usados apenas na avaliação final.

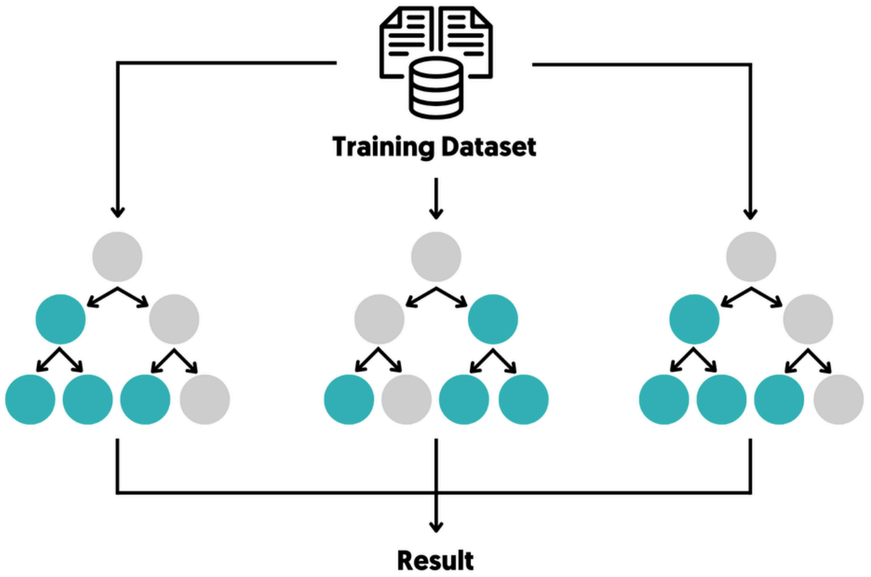
Classificação: Random Forest (Floresta Aleatória)

Pergunta: "É tóxico? Sim ou Não?" (Categórico).

Como funciona:

- Cria centenas de "Árvores de Decisão".
- Cada árvore vota na classe (Sim/Não).
- A decisão final é a maioria dos votos.

Vantagens: Robusto, lida bem com fingerprints (bits), fornece "importância das features".



Regressão

"Qual a solubilidade? 2.5 ou 4.1?" (Contínuo)

Regressão Linear:

- Tenta traçar uma reta (ou hiperplano) que passa o mais perto possível de todos os pontos.
- Equação: Propriedade = w_1*bit_1 + w_2*bit_2 + ... + constante.

Entendendo a Regressão Linear

Imagine um gráfico onde o eixo X é uma característica química e o eixo Y é a atividade.

O algoritmo busca a linha que minimiza a distância (erro) entre a linha e os pontos reais.

Limitação: Assume que a relação entre estrutura e propriedade é linear (o que nem sempre é verdade na química).

Como saber se o modelo é bom?

RMSE (Root Mean Squared Error):

- Mede o erro médio absoluto.
- Exemplo: Se prevemos solubilidade logS, um RMSE de 0.5 significa que erramos, em média, 0.5 unidades de log. (Quanto menor, melhor).

R^2 (Coeficiente de Determinação):

- Mede o quão bem o modelo explica a variação dos dados.
- $R^2 = 1.0$: Perfeito.
- $R^2 = 0.0$: O modelo é tão ruim quanto chutar a média.
- (Quanto maior, melhor).

Prática:

1. Carga: Importar o dataset com SMILES e Energia Experimental.
2. Limpeza: Remover sais e íons (SaltRemover) para isolar a molécula.
3. Features (X): Gerar Morgan Fingerprints (transformar desenho em vetor).
4. Split: Dividir dados em Treino e Teste.
5. Treino: Ajustar modelo de Regressão (prever números).
6. Avaliação: Calcular o erro médio (RMSE).
7. Visualização: Gerar gráfico de dispersão (Real vs. Predito).