Comparative Analysis of Drug Discovery Datasets

Abstract:

This study provides a comprehensive comparison of datasets commonly used in drug discovery resea

Introduction:
The quality and characteristics of datasets significantly impact the performance of machine learning m

Dataset Analysis:

1. ChEMBL Database:
- 2.3 million compounds
- 13,000 targets
- High-quality bioactivity data
- Standardized molecular representations

2. PubChem Database:
- 111 million compounds
- Diverse chemical space
- Multiple bioactivity assays
- Rich metadata

3. BindingDB:
- 1.6 million binding data points
- Protein-ligand interactions
- High-resolution structures
- Kinetic parameters

Evaluation Metrics:
- Data quality: Completeness and accuracy
- Chemical diversity: Structural variety
- Biological relevance: Target coverage
- Standardization: Format consistency

Conclusion:
ChEMBL provides the best balance of quality and size for most drug discovery applications, while Pub