Dataset

Data collection

![ChEMBL]

2255 approved SMILES

2255 unapproved SMILES

MOLFORMER Embedding

DMLC XGBoost

Similarity search

Data processing

5 most similar approved SMILES

Approved + Unapproved SMILES

5 most similar unaproved SMILES

Physiochemical feature extraction

RDKit

Custom dataset

System prompt

You are a chemist specializing in drug discovery and molecular modeling...
For each compound, you are provided with:
- Molecular descriptors and properties calculated via RDKit.
- The RDKit-derived descriptors and properties of the five most similar approved and five most similar unapproved small molecules (based on structural similarity).

Follow these steps:
1. Analyze the compound's physicochemical properties.
2. Compare the compound with its most similar approved and unapproved counterparts ....
3. Think step by step and use your knowledge of molecular approval criteria and the patterns ...
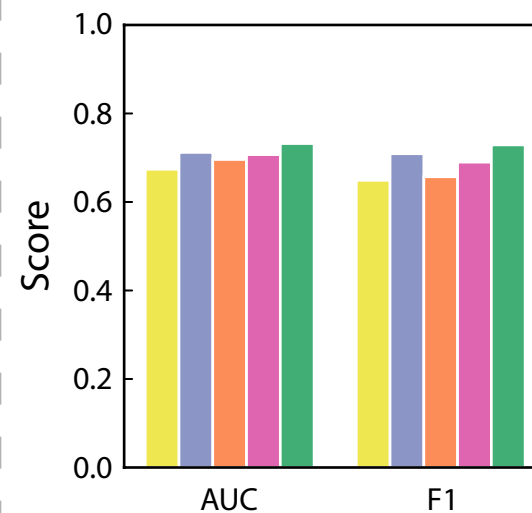
Based on your analysis, label the compound as either:
- approved
- unapproved

Then assign a confidence score between 0 and 1, where:
- A score near 1.0 indicates high confidence in your prediction (clear, strong evidence).
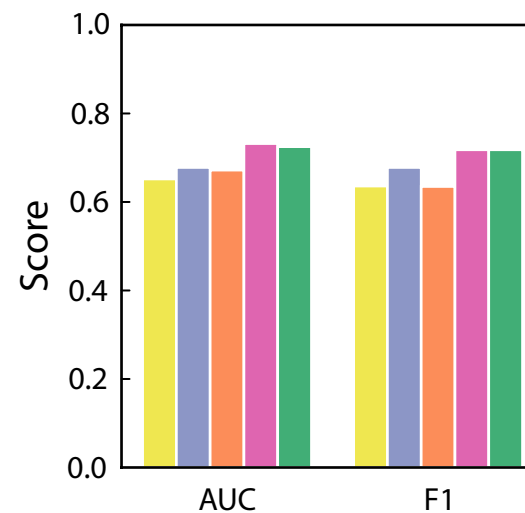- A score near 0.0 indicates high uncertainty (ambiguous, conflicting, or weak evidence).

Respond in the following format:
<think>
(Your step-by-step scientific reasoning)
</think>
<label>
(approved or unapproved)
</label>
<score>
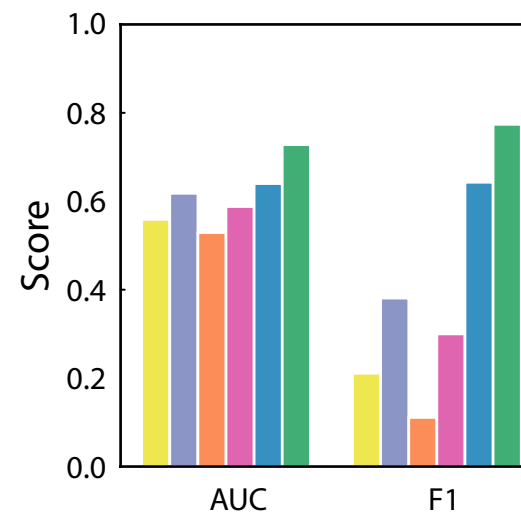(Numeric confidence score between 0 and 1)
</score>

Models assessment

Validation set / Test set / External set bar charts (Score vs AUC, F1)

Legend: Logistic regression, KNN, SVM, XGBoost, ChemAp, DrugReasoner

DrugReasoner development

- Correctness
- XML format
- Soft format compliance
- Interpretability
- Confidence-alignment

Reward

Llama-3.1-8B-Instruct

TRL

GRPO

DrugReasoner