Task 5: Hit Identification – Compound Mining

Overview

The objective of Task 5 was to identify and prioritize potential small-molecule inhibitors for DNA gyrase through systematic compound mining, physicochemical filtering, and hit shortlisting. Public bioactivity data were retrieved from ChEMBL and processed using Python and RDKit to generate a curated, drug-like compound dataset. Classical drug-likeness rules were applied to prioritize viable candidates, resulting in a ranked shortlist of hits supported by activity data, molecular descriptors, and structural similarity analysis.

Compound Retrieval and Dataset Curation

DNA gyrase was queried in ChEMBL, and associated IC50 activity records were extracted together with canonical SMILES. The raw dataset was converted into a pandas DataFrame and curated by removing missing values, validating SMILES using RDKit, and deduplicating compounds based on canonical SMILES. Experimental IC50 values were normalized by conversion to pIC50 to enable consistent activity comparison.

The curated dataset contained ChEMBL compound identifiers, canonical SMILES, standardized activity values, and pIC50, providing a clean foundation for downstream cheminformatics analysis.

Descriptor Generation and Drug-Likeness Filtering

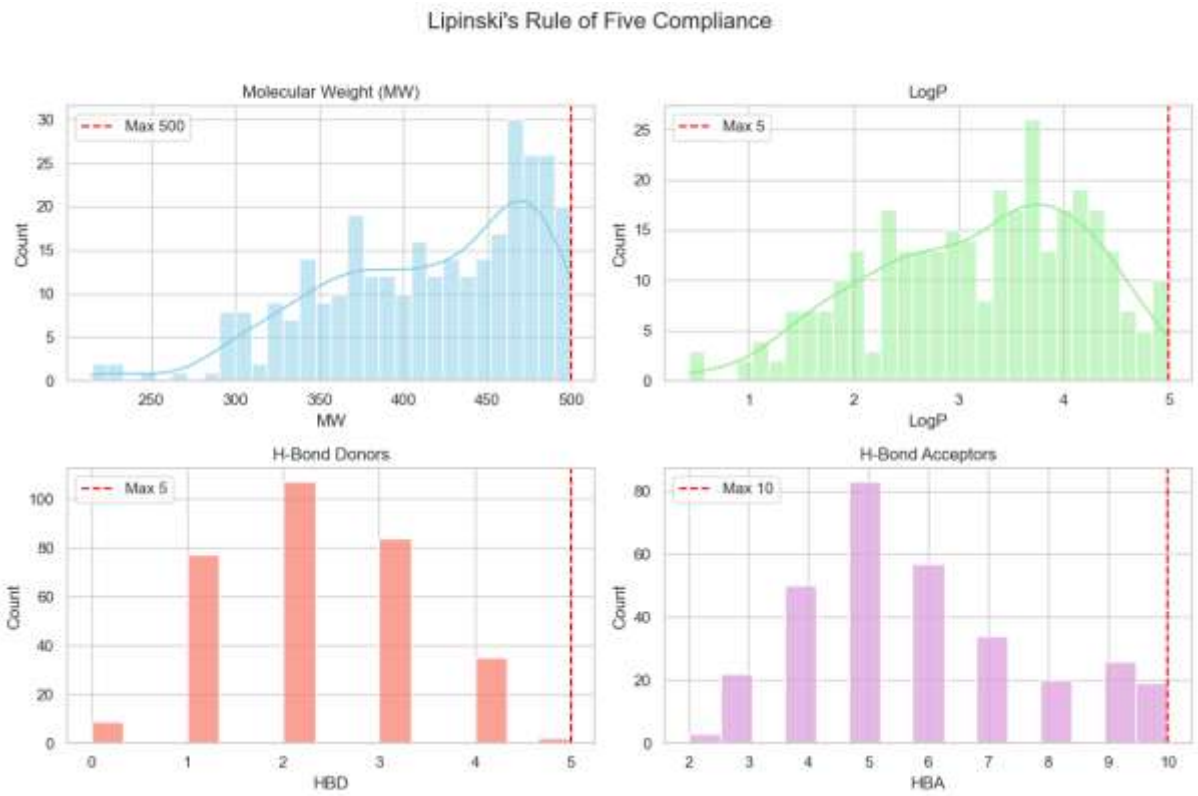
RDKit was used to compute key physicochemical descriptors for each compound, including:

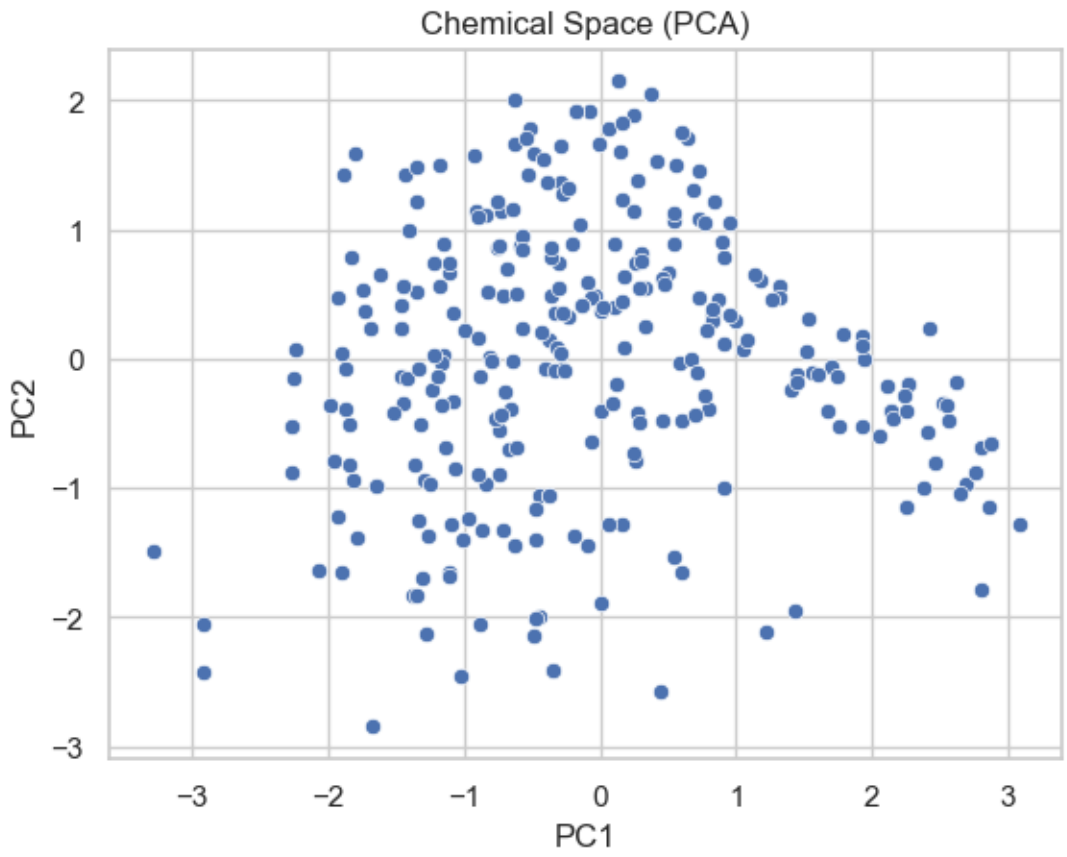
- Molecular Weight (MW)
- LogP
- Hydrogen Bond Donors (HBD)
- Hydrogen Bond Acceptors (HBA)
- Topological Polar Surface Area (TPSA)
- Number of Rotatable Bonds

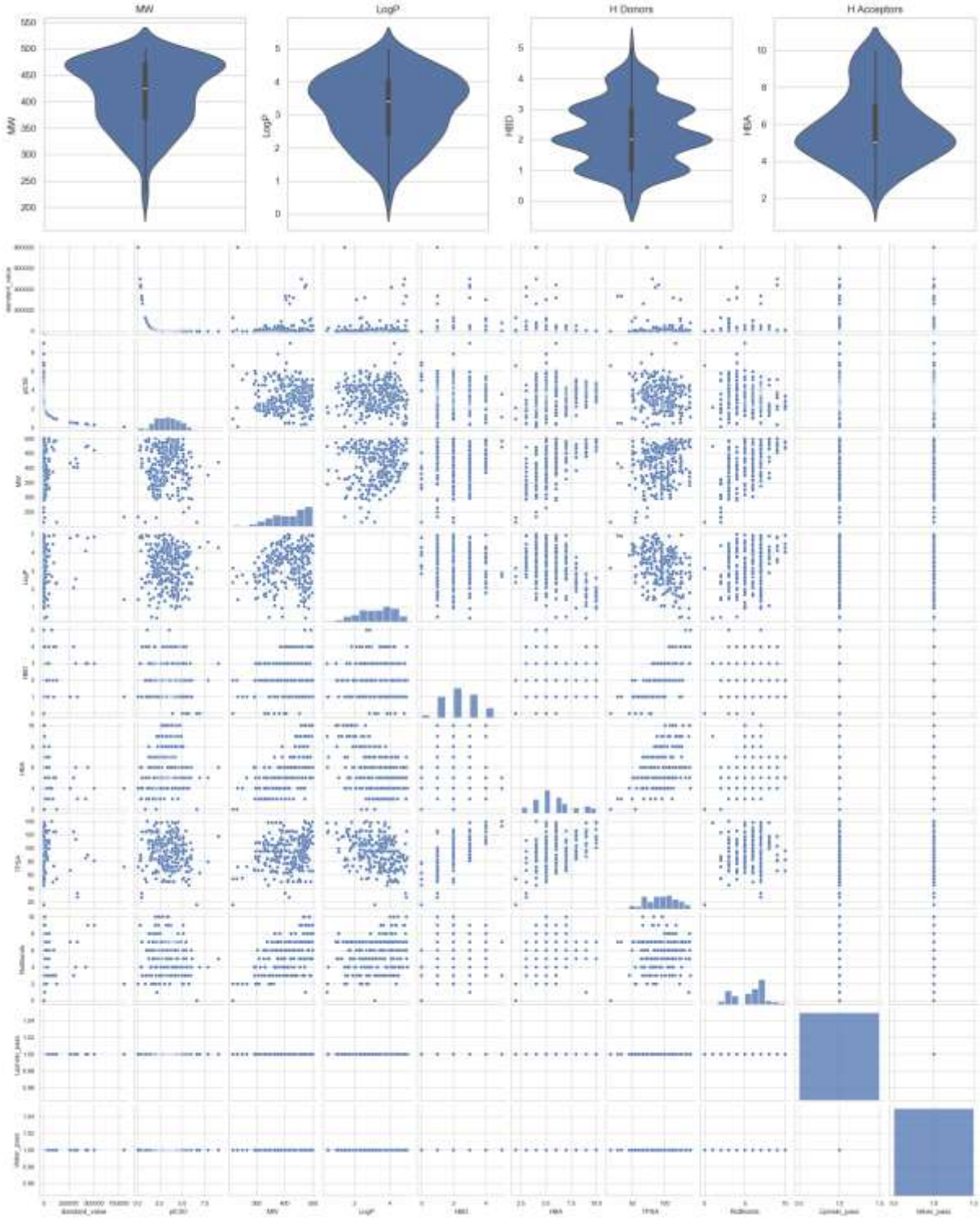
Drug-likeness was assessed using Lipinski's Rule of Five (MW $\leq$ 500, LogP $\leq$ 5, HBD $\leq$ 5, HBA $\leq$ 10) and Veber criteria (TPSA $\leq$ 140 Å^2, rotatable bonds $\leq$ 10). Only compounds passing both Lipinski and Veber filters were retained, ensuring that shortlisted molecules exhibit physicochemical profiles compatible with oral bioavailability.

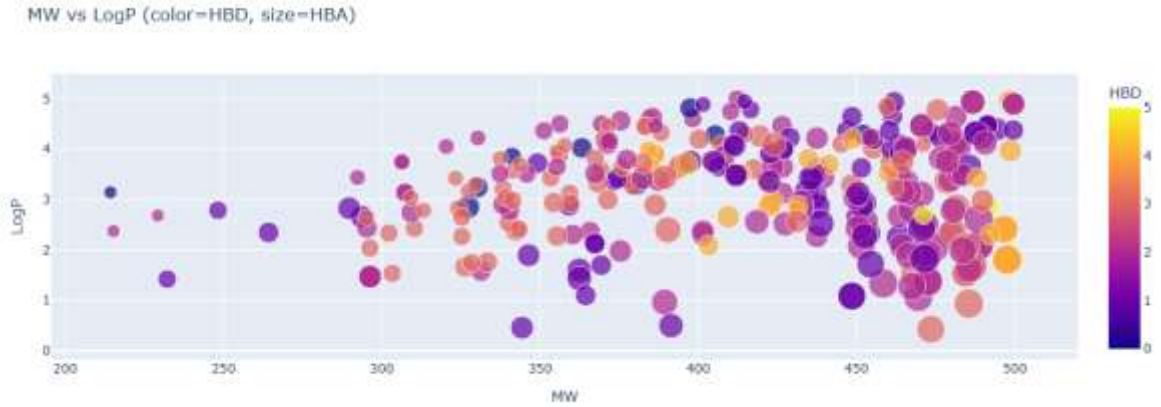
Exploratory Analysis and Chemical Space

The filtered dataset was explored using descriptor histograms, violin plots, pairwise relationships, and interactive MW vs LogP scatter plots. Principal Component Analysis (PCA) was performed to visualize chemical space and assess compound diversity. These analyses confirmed that the retained compounds occupy a coherent and drug-like chemical space, with reasonable distributions across molecular weight, lipophilicity, and polarity.


Lipinski's Rule of Five Compliance
Molecular Weight (MW)
Max 500
Count
MW
LogP
Max 5
Count
LogP
H-Bond Donors
Max 5
Count
HBD
H-Bond Acceptors
Max 10
Count
HBA


Chemical Space (PCA)
PC2
PC1


MW
LogP
H Donors
H Acceptors
MW
LogP
HBD
HBA



Hit Shortlisting and Ranking

Compounds passing Lipinski and Veber filters were ranked by pIC50, and the top 20 candidates were shortlisted as preliminary hits. The highest-ranked compounds included:

- **CHEMBL3094348** (pIC50 = 9.00, MW = 419.5, LogP = 4.26)
- **CHEMBL3094347** (pIC50 = 7.85, MW = 375.5, LogP = 4.56)
- **CHEMBL515530** (pIC50 = 6.90, MW = 405.4, LogP = 4.28)
- **CHEMBL463832** (pIC50 = 6.60, MW = 214.3, LogP = 3.14)
- **CHEMBL390983** (pIC50 = 6.08, MW = 401.4, LogP = 2.31)

Each shortlisted compound satisfies Lipinski and Veber criteria and demonstrates favorable predicted potency. Automated rationale strings were generated for each hit, summarizing pIC50, drug-likeness compliance, and key physicochemical properties.

Structural Similarity Analysis

To contextualize novelty, Tanimoto similarity was calculated between shortlisted hits and ciprofloxacin as a reference DNA gyrase inhibitor. All shortlisted compounds exhibited low similarity scores (typically ~0.06–0.08), indicating that the identified hits represent chemically distinct scaffolds rather than close analogs of fluoroquinolones. This suggests potential novelty while retaining predicted bioactivity.

Outcome

Task 5 produced a curated, descriptor-rich compound dataset and a ranked shortlist of drug-like DNA gyrase inhibitors supported by activity normalization, physicochemical filtering, chemical space analysis, and structural similarity assessment. The final output includes a CSV-exported hit table containing ChEMBL IDs, SMILES, pIC50, molecular descriptors, and automated rationale for each candidate. This workflow demonstrates an end-to-end hit identification pipeline and provides a strong foundation for subsequent QSAR modeling or structure-based validation.