Result1: Read with sep='\t'

First 6 rows (raw):

Compound Name Class PCID Water Solubility Caco2 Permeability Intestinal Absorption (%) Skin Permeability P-gp Substrate CYP3A4 Substrate CYP Inhibitors BBB Permeability Total Clearance AMES Toxicity hERG I Inhibitor hERG II Inhibitor Hepatotoxicity Unnamed: 16 Unnamed: 17 Unnamed: 18

Acetyl tyrosine ethyl ester Amino acid 2004 -2.089 0.453 82.974 -2.922 No No NaN -0.249 0.645 Yes No No No NaN NaN NaN

N(alpha)-Benzyloxycarbonyl-Lleucine Amino acid 74840 -2.701 0.946 91.584 -2.728 Yes No NaN -0.131 0.598 No No No No NaN NaN NaN

8-Desoxygartanin Xanthone 392450 -4.075 1.139 90.497 -2.735 Yes Yes CYP1A2;CYP2C19;CYP2C9 -0.87 0.329 Yes No Yes Yes NaN NaN NaN

Lophophorine Alkaloids 442315 -1.499 1.876 92.430 -2.647 No Yes CYP2D6 0.262 0.926 No No No No NaN NaN NaN

L-trans-5-Hydroxy-2piperidinecarboxylic acid Piperidines 151730 -0.102 0.492 70.428 -2.735 No No NaN -0.75 0.666 No No No No NaN NaN NaN

Candimine Alkaloids 441588 -3.395 0.806 64.066 -3.235 Yes Yes NaN -0.286 0.687 Yes No No Yes NaN NaN NaN

Detected numeric-like columns (7): ['PCID', 'Water Solubility', 'Caco2 Permeability', 'Intestinal Absorption (%)', 'Skin Permeability', 'BBB Permeability', 'Total Clearance']

Detected non-numeric columns (12): ['Compound', 'Class', 'P-gp Substrate', 'CYP3A4 Substrate', 'CYP Inhibitors', 'AMES Toxicity', 'hERG I Inhibitor', 'hERG II Inhibitor', 'Hepatotoxicity', 'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18']

Example rows with NaNs in numeric columns (inspect for parsing issues):

Compound PCID Water Solubility Caco2 Permeability Intestinal Absorption (%) Skin Permeability BBB Permeability Total Clearance

8-Deoxy-11 NaN 5260103.0 -2.645000e+00 0.524 68.497 NaN 0.013

2-Hexyl-1 NaN 74361.0 -2.003000e+00 1.707 93.821 NaN 0.580

7(14)-Bisabolene-2 10.0 NaN NaN 73816843.000 -1.942 NaN NaN

1-Methoxy-1-(2 NaN NaN 1.317512e+08 -1.751 1.271 NaN NaN

3-Oxo-12 NaN 14707579.0 -3.201000e+00 1.260 100.000 NaN NaN

Numeric matrix ready. Shape: (32, 7)

Result 2:

R2 (test): -0.652

RMSE (test): 0.436

Saved ranked CSVs: ranked\_predictions\_export.csv top\_candidates.csv

Saved toxicity-filtered candidates: safe\_candidates.csv

Saved PCA plot: pca\_plot.png

/usr/local/lib/python3.12/dist-packages/sklearn/manifold/\_t\_sne.py:1164: FutureWarning: 'n\_iter' was renamed to 'max\_iter' in version 1.5 and will be removed in 1.7.

warnings.warn(

Saved t-SNE plot: tsne\_plot.png

Top 10 predicted compounds:

Compound Predicted\_Score

1 6-Hydroxypentadecanedioic acid 0.7368

2 1-Methoxy-1-(2 0.7029

3 5-(2-Methylpropyl)tetrahydro2-oxo-3-furancarboxylic acid 0.6543

4 1alpha,2alpha,4betaH,6alpha,8R)-p-Menthane-2,6,8,9-tetrol, 0.5083

5 Schizonepetoside E 0.4349

6 7(14)-Bisabolene-2 0.3625

7 Di-4-coumaroylputrescine 0.2201

8 4-(3-Pyridyl)-3-butenoic acid 0.2062

9 Methyl o-methoxyhippuric acid 0.1379

10 8-Deoxy-11 0.1344

Outputs saved in: /content

Result 3:

Moschamine found in ranked predictions

Rank: 1

Predicted\_Score: 0.0855

Moschamine descriptors from original dataset:

25

Compound Moschamine

Class Amines

PCID 5969616

Water Solubility -3.739

Caco2 Permeability -0.1

Intestinal Absorption (%) 78.363

Skin Permeability -2.737

P-gp Substrate Yes

CYP3A4 Substrate Yes

CYP Inhibitors CYP1A2;CYP2C19;CYP2C9

BBB Permeability -1.022

Total Clearance 0.462

AMES Toxicity No

hERG I Inhibitor No

hERG II Inhibitor Yes

Hepatotoxicity Yes

Unnamed: 16 NaN

Unnamed: 17 NaN

Unnamed: 18 NaN

Saved moschamine\_and\_top10.csv in /content

Result 4:

Proxy built using numeric features: ['PCID', 'Caco2 Permeability', 'Intestinal Absorption (%)', 'Water Solubility', 'Skin Permeability', 'BBB Permeability']

Toxicity-like columns considered for penalty: ['AMES Toxicity', 'hERG I Inhibitor', 'hERG II Inhibitor', 'Hepatotoxicity']

Saved per-compound errors to per\_compound\_errors.csv

Compounds evaluated: 32

Overall MSE: 0.188337

Overall RMSE: 0.433978

Top 10 worst-predicted compounds by RMSE\_i:

Compound Interaction\_Score\_true Predicted\_Score Error Squared\_Error RMSE\_i

3-Oxo-12 1.000000 0.1271 -0.872900 0.761954 0.872900

Lophophorine 0.686652 0.0449 -0.641752 0.411846 0.641752

N(alpha)-Benzyloxycarbonyl-Lleucine 0.655247 0.0291 -0.626147 0.392060 0.626147

Kolanone 0.705245 0.0855 -0.619745 0.384084 0.619745

Diethyl (2R 0.632404 0.0392 -0.593204 0.351891 0.593204

L-trans-5-Hydroxy-2piperidinecarboxylic acid 0.606759 0.0302 -0.576559 0.332420 0.576559

5-Oxoavermectin''1b'' aglycone 0.623605 0.0788 -0.544805 0.296812 0.544805

Suberylglycine 0.600681 0.0675 -0.533181 0.284282 0.533181

N-Acetyl-DL-methionine 0.643792 0.1265 -0.517292 0.267591 0.517292

1-O-Sinapoyl-β-D-glucose 0.569126 0.0559 -0.513226 0.263401 0.513226

Result 5:

Compound Interaction\_Score\_true Predicted\_Score Error Squared\_Error RMSE\_i

Moschamine 0.193985 0.0855 -0.108485 0.011769 0.108485

Result 6:

Compound Interaction\_Score\_true Predicted\_Score Error Squared\_Error RMSE\_i RMSE\_rank

Moschamine 0.193985 0.0855 -0.108485 0.011769 0.108485 29

Compound Interaction\_Score\_true Predicted\_Score Error Squared\_Error RMSE\_i Rank\_by\_score

Moschamine 0.193985 0.0855 -0.108485 0.011769 0.108485 1