Synergizing Pharmacophore Screening and Machine Learning: *Identification of Novel Cathepsin L Inhibitors*

Elisabetta Grazia Tomarchio[1,2], Verena Battisti[3], Oliver Wieder[3], Antonio Rescifina[2], Thierry Langer[3]

[1]*Department of Drug and Health Sciences, University of Catania, Viale A. Doria 6, 95125, Catania, Italy*

[2]*Department of Biomedical and Biotechnological Sciences, University of Catania, Via Santa Sofia 97, 95123 Catania, Italy*

[3]*Department of Pharmaceutical Chemistry, Althanstrasse 14 (UZA II), 1090 Vienna, Austria*

The high active-site homology across cysteine cathepsins, particularly within the S2 subsite, makes selective inhibition of Cathepsin L (CTSL) a challenging molecular design problem.[1] CTSL is implicated in tumor metastasis and therapy resistance, highlighting the need for selective chemical probes.[2] We developed an integrated in silico workflow combining structure-based pharmacophore modeling and machine learning (ML) to identify novel CTSL inhibitors. High-resolution CTSL crystal structures were used to generate pharmacophore models capturing key S2 subsite interactions and anchoring to catalytic Cys25.[3] After rigorous validation, the models were applied to screen multi-million compound libraries. Hits were further prioritized using a QSAR regression framework built on a curated dataset of 1,218 inhibitors (pActivity range: 2–10). Molecular features were encoded using ECFP4 fingerprints and physicochemical descriptors. To assess generalization robustness, three data-splitting strategies (random, Bemis–Murcko scaffold-based, and UMAP-cluster-based splits) were evaluated. For each split, a stacked ensemble combining Random Forest and XGBoost regressors was trained. Final prioritization relied on consensus predictions across all split-specific models to mitigate split-dependent bias and enhance scaffold diversity. The optimized models demonstrated robust predictive performance for a structurally diverse CTSL dataset (R^2 = 0.662 ± 0.03; Pearson r = 0.81; Q^2_LOO = 0.565; MAE = 0.595 ± 0.03; RMSE = 0.773 ± 0.04). This workflow enables efficient exploration of large chemical spaces and provides a reproducible strategy for prioritizing selective CTSL inhibitor candidates for experimental validation.

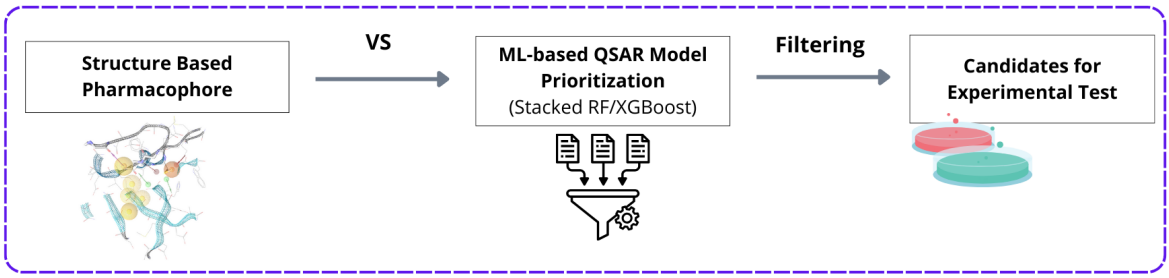

Figure 1. Schematic representation of the workflow used in this work.

Bibliography:

[1] Abdelaziz, R. F.; Hussein, A. M.; Kotob, M. H.; Weiss, C.; Chelminski, K.; Stojanovic, T.; Studenik, C. R.; Aufy, M. *IJMS* (2023), *24*, 17106.
[2] Sudhan, D. R.; Siemann, D. W. *Pharmacology & Therapeutics* (2015), *155*, 105–116.
[3] Shenoy, R. T.; Sivaraman. *Journal of Structural Biology* (2011), *173*, 14–19.