ECO-CONSCIOUS ANTIPARASITIC DISCOVERY: A UNIFIED ML-DRIVEN AND ECOTOXICOLOGICAL PRIORITIZATION STRATEGY

D. Aiello[1], L. Bertarini[1,2,3], R. Karki[4,5], S. Gul[4,5], F. Pellati[1], M. Tonelli[6], E. Uliassi[7], C. Borsari[8], V. Tudino[9], S. Gemma[9], T. Calogeropoulou[10] and M. P. Costi[1]

[1]Department of Life Sciences, University of Modena and Reggio Emilia, Via G. Campi 103, 41125 Modena, Italy.[2] Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Via Campi 287, 41125 Modena, Italy. [3] Clinical and Experimental Medicine PhD Program. [4] Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Discovery Research ScreeningPort, D-22525 Hamburg, Germany. [5] Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany. [6] Department of Pharmacy, University of Genoa, Viale Benedetto XV 3, 16132 Genoa, Italy.[7]Department of Pharmacy and Biotechnology, Alma Mater Studiorum—University of Bologna, Via Belmeloro 6, I-40126 Bologna, Italy, [8]Department of Pharmaceutical Sciences, University of Milan, Via Mangiagalli 25, 20133, Milan, Italy, [9]TheraFood Research, Department of Biotechnology, Chemistry and Pharmacy, University of Siena, 53100 Siena, Italy, [10]Institute of Chemical Biology, National Hellenic Research Foundation, 48 Vassileos Constantinou Avenue, Athens 11635, Greece

Introduction

Neglected parasitic diseases continue to impose a substantial global health burden, yet early-stage antiparasitic discovery rarely integrates environmental safety considerations. Traditional hit prioritization focuses on biochemical potency, cytotoxicity, and pharmacokinetic properties, overlooking potential ecotoxicological risks associated with new chemical entities. To address this gap, we assembled a unified dataset of antiparasitic compounds active against *Babesia*, *Leishmania*, *Schistosoma*, and *Trypanosoma* spp., derived from peer-reviewed studies published between 2019 and 2024. Each compound was curated from the literature by extracting structural information, activity data (IC_{50} and Ki), phenotypic potency (IC_{50} < 10 µM required), selectivity information, cytotoxicity profiles, and available in-vivo evidence. To complement biological data, ecotoxicological parameters—BCF, IGC50, LC50DM, and LC50FM—were predicted using ADMETlab 3.0. Integrating these environmental descriptors with ADMET and drug-likeness properties enabled the development. Building upon this integrated biological and ecotoxicological dataset, the study pursued two main goals:

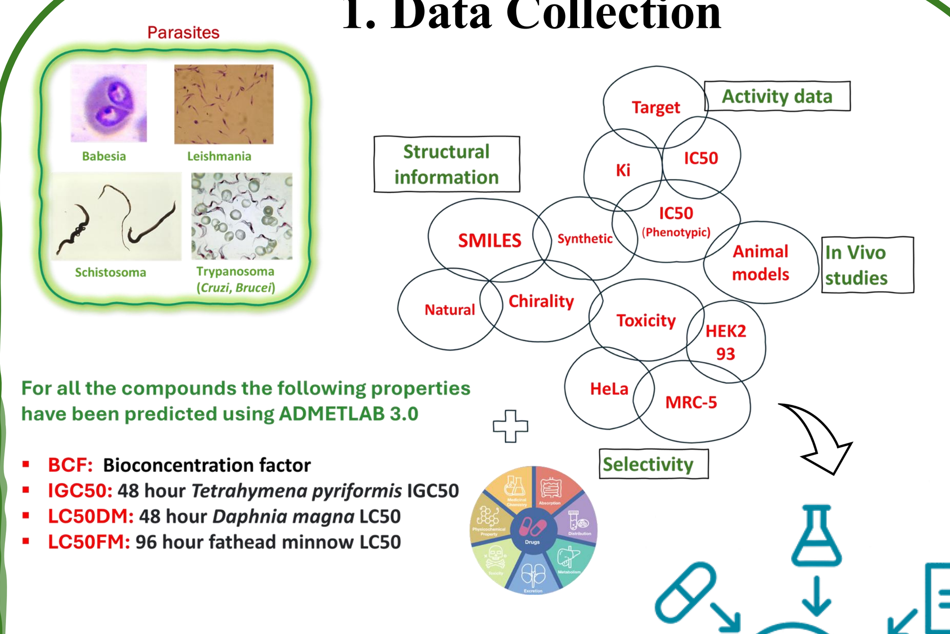
First Goal

First, we sought to determine whether incorporating environmental toxicity endpoints into early screening genuinely reshapes compound prioritization—potentially altering which molecules would be selected as hits under conventional criteria.

Second Goal

Second, we aimed to identify **environmentally favourable chemotypes** within the antiparasitic chemical space, providing safer and more sustainable starting points for future drug-discovery efforts.

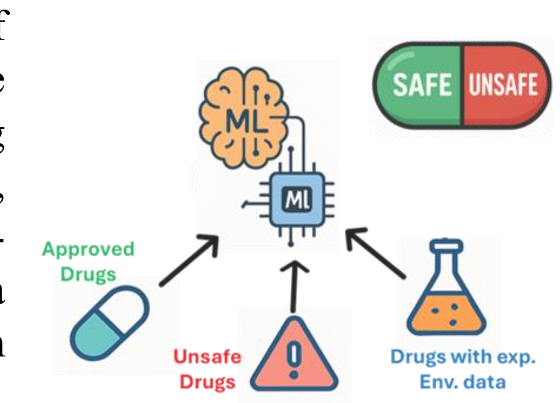
1. Data Collection



Virtual Chemotheca Data Collection and Management

2. Machine Learning Classifier Training and Performance

To support early detection of potentially unsafe compounds, we built a curated dataset integrating FDA-approved drugs (SAFE), withdrawn drugs (UNSAFE), phase-II failures from ChEMBL, and a small set of molecules with experimental ecotoxicity data, yielding **1464 compounds** spanning diverse chemical space.

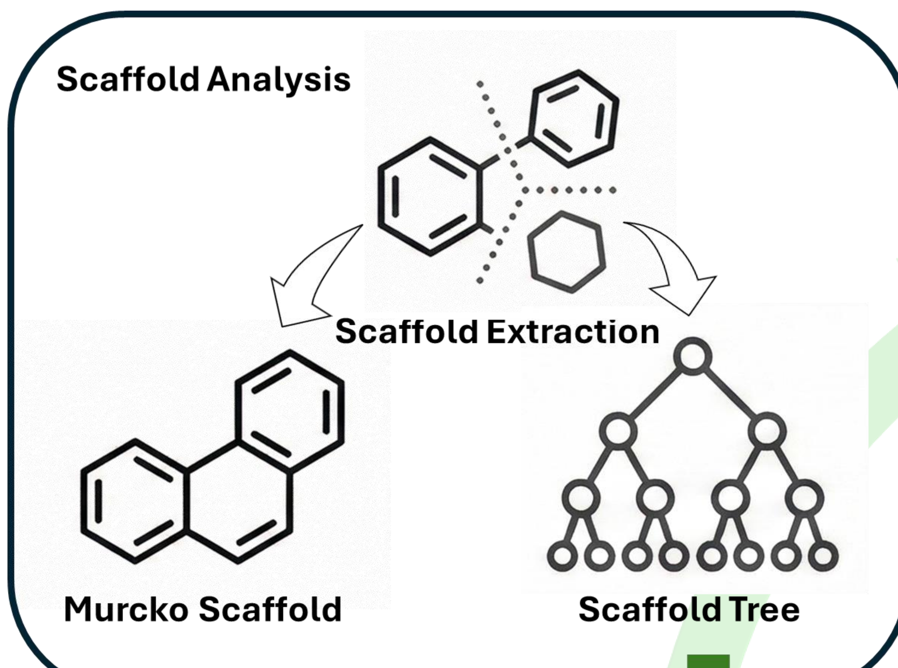


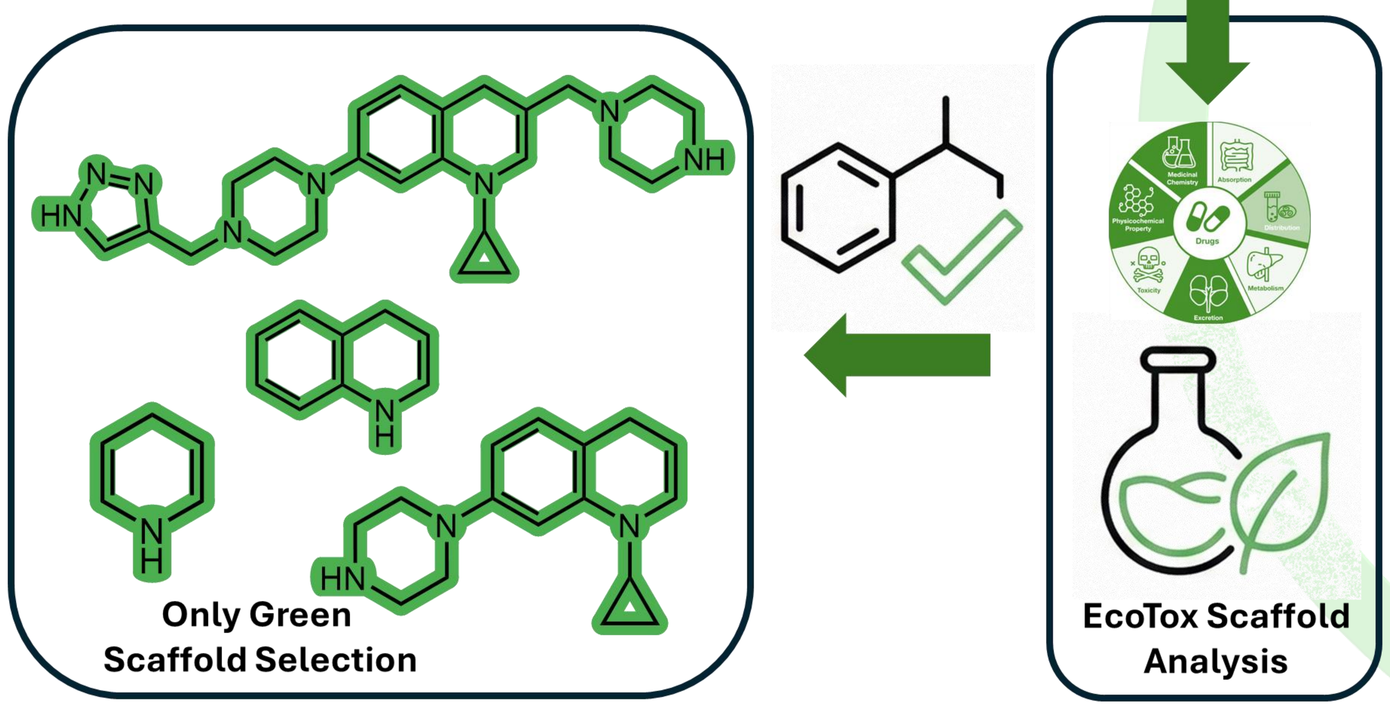
Classifier	ROC AUC	Bal. Accuracy	PR AUC	MCC
XGBoost (XGB)	**0.851**	0.773	0.837	**0.553**
AdaBoost (ADA)	0.744	0.676	0.725	0.357
Gradient Boosting (GB)	0.759	0.688	0.724	0.357
Extra Trees (ET)	0.787	0.709	0.755	0.410
CART	0.739	0.665	0.726	0.318
Random Forest (RF)	0.810	0.735	0.783	0.464

Across all tested algorithms, **XGBoost** provided the most robust SAFE/UNSAFE classification, with a ROC AUC of 0.851, balanced accuracy of 0.773, and MCC of 0.553. Its strong ROC and precision–recall profiles made XGBoost the optimal model for subsequent scoring and hit-prioritization steps.

4. Scaffold Analysis and Green Scaffold Selection

To understand how environmental constraints shape the antiparasitic chemical space, we performed a scaffold-centered evaluation of all compounds ranked with the GreenDrugScore. Bemis–Murcko scaffolds were first extracted to capture the core structural frameworks, followed by a hierarchical scaffold tree to resolve substructures and recurring motifs.





By mapping these scaffolds against GDS performance and ecotoxicity classes, we identified structural families consistently associated with favourable environmental profiles. This approach enabled the recognition of **green chemotypes**—core motifs that combine potent antiparasitic activity with minimal predicted ecological impact—providing sustainable starting points for future hit-optimization and drug-design efforts.

3. GreenDrugScore and Ranking

To integrate biological safety, environmental impact, and drug-likeness into a single prioritization metric, we developed the GreenDrugScore (GDS). For every compound, **ADMETscore**, **ECOscore**, and **DLscore** were calculated and used to generate a global ranking. Each component is computed as a weighted average of its endpoints, with weights reflecting endpoint entropy, predictive accuracy, and biological or regulatory relevance, so that the most informative descriptors drive the final score.

$$\text{ADMETscore} = \frac{\sum w_i \cdot q_i}{\sum w_i}$$

hERG
DILI
AMES
CYP3A4
P-gp

+

$$\text{ECOscore} = \frac{\sum w_i \cdot q_i}{\sum w_i}$$

BCF
IGC50
LC50DM
LC50FM

+

$$\text{DLscore} = \frac{\sum w_i^{DL} \cdot q_i^{DL}}{\sum w_i^{DL}}$$

ClogP
QED
TPSA
Lipinski

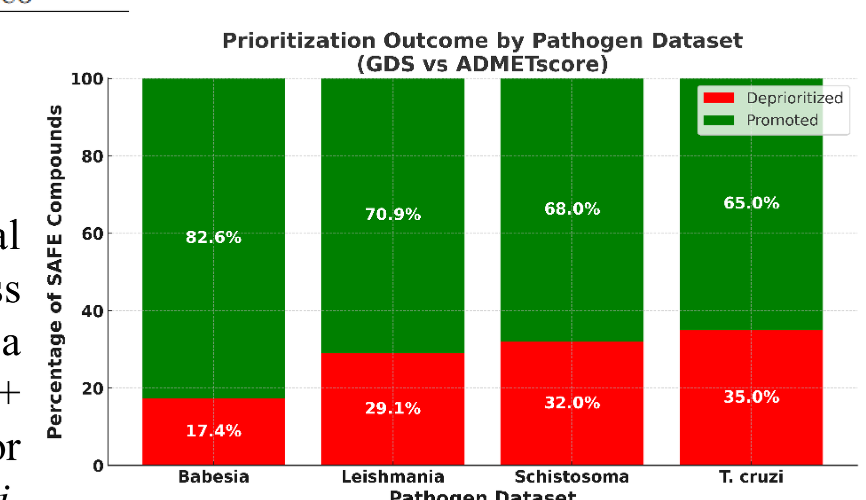
Weights
w1: endpoint entropy
w2: model accuracy (e.g. AUC)
w3: biological or regulatory importance

Total endpoints
ADMETscore: 21
ECOscore: 4
Dlscore: 11

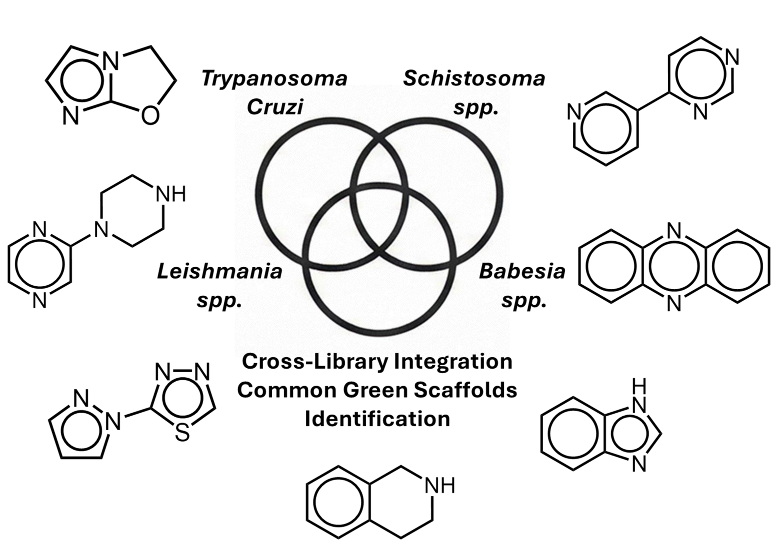
ADME
Eco-Tox
Drug-Likeness

$$\text{GDS}_{\text{2-block}} = \frac{w_{\text{ADMET}} \cdot \text{ADMET} + w_{\text{ECO}} \cdot \text{ECO}}{w_{\text{ADMET}} + w_{\text{ECO}}}$$

The framework includes 21 ADMET endpoints, 4 ecotoxicological endpoints (BCF, IGC50, LC50DM, LC50FM), and 11 drug-likeness descriptors (including QED, clogP, TPSA, Lipinski). When moving from a ranking based solely on ADMETscore to the 2-block GDS (ADMET + ECO), a fraction of SAFE compounds was deprioritized: about 17% for *Babesia*, 29% for *Leishmania*, 32% for *Schistosoma* and 35% for *T. cruzi*, with the remaining SAFE hits being promoted.



5. Conclusions



Across the four parasite-focused libraries, we identified a total of 241 fully green scaffolds, distributed as follows: 112 unique to *Leishmania*, 88 to *T. cruzi*, 28 to *Schistosoma*, and 13 to *Babesia*. Among these, **38 scaffolds were shared across multiple parasites**, representing conserved eco-friendly chemotypes with potential broad applicability.

The remaining scaffolds were parasite-specific, highlighting distinct structural preferences within each biological system.

Overall, this study enabled the identification of a substantial number of scaffolds with optimal environmental profiles and demonstrated that incorporating ecotoxicological endpoints **significantly influences early hit selection**, reshaping prioritization toward more sustainable antiparasitic candidates

REFERENCES

[1] J.C. Semenza, S. Paz, Climate change and infectious disease in Europe: Impact, projection and adaptation, Lancet Reg. Health Eur. 9 (2021) 100230. https://doi.org/10.1016/j.lanepe.2021.100230. [2] Aiello, D.; Bertarini, L.; Karki, R.; Gul, S.; Pellati, F.; Tonelli, M.; Costi, M. P. Leveraging Ecotoxicity Parameters and Machine Learning to Redefine the Drug Discovery Pipeline. J. Pharm. Anal. 2025, under review.[3] L. Fu, S. Shi, J. Yi, N. Wang, Y. He, Z. Wu, J. Peng, Y. Deng, W. Wang, C. Wu, A. Lyu, X. Zeng, W. Zhao, T. Hou, D. Cao, ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support, Nucleic Acids Res. 52 (2024) W422–W431. https://doi.org/10.1093/nar/gkae236.