How to Select the Right Biological Data to Enable Effective Drug Discovery

Aleksandar Pavic, University of Belgrade

Despite substantial and sustained investment by the pharmaceutical industry in the development of novel anticancer therapies, the discovery of new antimicrobial agents, particularly those targeting fungal pathogens, remains critically underdeveloped. This gap is especially concerning in the context of rising antifungal resistance, limited therapeutic options, and the high mortality associated with invasive fungal infections. On the other side, antimicrobial drug resistance and the incidence of lethal infection-induced outcomes are being rapidly increased.

Artificial intelligence (AI) and machine learning (ML) offer unprecedented opportunities to accelerate antimicrobial drug discovery by enabling the identification of compounds with entirely novel mechanisms of action and by predicting key properties such as biological activity, potency, biosafety, and in vivo efficacy. However, the success of AI-and ML-driven drug discovery critically depends not only on algorithmic sophistication but, more fundamentally, on the quality, relevance, and biological meaning of the data used for model training and validation.

Most available antimicrobial activity data are generated using standardized EUCAST or CLSI protocols, while toxicity data are predominantly derived from in vitro assays on immortalized cell lines, and to a much lesser extent from in vivo models such as zebrafish (Danio rerio), Galleria mellonella, or mammalian systems (mice, rat, rabbit, primates). Although these datasets are increasingly compliant with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles - ensuring transparency, accessibility, and reusability - FAIR compliance alone does not guarantee biological relevance or translational value.
A central challenge in AI-assisted antimicrobial discovery is therefore not data availability, but data selection: determining which biological datasets are truly informative, predictive, and mechanistically meaningful for computational modeling. This lecture will address key questions in data-driven drug discovery: Are all FAIR datasets equally relevant? How can we identify and prioritize biologically and clinically meaningful data for AI/ML-based analysis? Which experimental models best capture host–pathogen interactions, toxicity, and efficacy in a way that improves translational success?

Finally, this lecture will critically evaluate real-world outcomes of AI-driven drug discovery by highlighting antimicrobial and non-antimicrobial drug candidates identified through bioinformatics, cheminformatics, and AI/ML approaches that have progressed to clinical trials or received FDA and EMA approval. By integrating biological insight with computational rigor, this talk aims to provide a practical framework for selecting the “right” data to enable the development of safe, effective, and next-generation antimicrobia.