Enhanced Molecule Selection via Active Learning in Lead Optimization for Drug Discovery

Pablo Mas[1,2], Rodolphe Vuileumier[1], Marc Bianciotto[2]

[1]*PASTEUR, Département de chimie, École Normale Supérieure, PSL (Paris Sciences et Lettres) Université, Sorbonne Université, CNRS, 75005 Paris, France.*

[2]*Molecular Design Sciences - Integrated Drug Discovery, Sanofi R&D, 94400 Vitry-sur-Seine, France*

Active learning, a subset of machine learning, is transforming drug discovery by enabling the efficient analysis and prediction of biological activities and properties of compounds. This approach is particularly advantageous for quantitative structure-activity relationships (QSAR) modeling, as it significantly reduces the resources required to develop accurate predictive models. By strategically selecting the most informative samples, active learning allows for model training with fewer labeled data points, thus expediting the exploration of chemical space. Traditionally, active learning has been predominantly applied during the screening phase of drug discovery. However, this study explores the potential benefits of implementing active learning during the lead optimization phase.

Lead optimization is a critical stage in drug discovery, focusing on the refinement of lead compounds to enhance their efficacy, selectivity, and safety. The objective is to optimize drug-like properties, such as potency and ADMET (absorption, distribution, metabolism, excretion, and toxicity) profiles, to produce viable drug candidates for clinical trials. While the design of new molecules has historically been led by medicinal chemists, generative artificial intelligence (AI) has introduced new methodologies for generating large sets of molecules under multi-parametric optimization. Given the physical synthesis limitations faced by chemists, only a few tens of molecules per month can be synthesized in typical projects which makes the selection of molecules for synthesis a pivotal decision.

This work introduces a novel application of active learning principles in lead optimization, employing a retrospective analysis of internal drug discovery projects. We compare what happened during a lead-optimization project with hypothetical scenarios where molecule selection is guided by active learning strategies. Unlike virtual screening, our approach incorporates a temporal dimension, acknowledging the continuous generation of new molecular candidates throughout the project, as well as the multi-parametric optimization challenges, including efficacy and ADMET properties. The case study presented spans a three-year project timeline, with a focus on optimizing 12 distinct properties, demonstrating the strategic advantage of using active learning approaches in the later stages of drug discovery.

Our findings suggest that active learning can play a transformative role in lead optimization, potentially leading to more efficient and effective drug development processes. By integrating active learning into this phase, we aim to streamline the selection of new drug candidates, thereby enhancing the overall productivity and success rates of drug discovery programs.

Pablo Mas and Marc Bianciotto are Sanofi employees and may hold shares and/or stock options in the company. Rodolphe Vuilleumier has nothing to disclose.