How to boost omics biology research using generative artificial intelligence tools while avoiding the pitfalls of hallucinations

Thomas Burger*[1]

[1]CEA Grenoble (BIG, Biologie à grande Echelle, EDyP) – INSERM U1038, Université Grenoble Alpes – 17 rue des Martyrs 38054 Grenoble Cedex 9, France

Abstract

Generative artificial intelligence can be used to generate realistic new data, even for complex real-word processes which cannot be exhaustively formally modelled. In these cases, the model is simply learnt from pre-existing data. Generative artificial intelligence is therefore expected to be a game-changer in omics research, where new data collection is hampered by considerable experimental constraints. However, generative artificial intelligence tools can also "hallucinate", i.e., generate data which are too original to be realistic. In contrast to a classical machine learning-based prediction, where discrepancies with respect to the expected answer can be objectively measured, it is not easy to delimit the creativity/hallucination continuum. The difficulty is even greater in domains like molecular biology, that remain partly unexplored, and where erroneous inferences could have devastating consequences.
In this context, I propose a risk-mitigation policy that extrapolates on the principles that motivate the use of the false discovery rate control in omics data analysis and biomarker screening: By drawing a comparison between classical Type I errors and undetected hallucinations, it is possible to distinguish riskier and safer use-cases, depending on how undetected hallucinations translates into false biological discoveries.

Based on this principle, it possible to explore various families of use-cases where the full potential of generative methods applied to omics biology research can be unleashed. We present 3 such families, each illustrated with several use-cases: those where generative artificial intelligence is used for pre-screening biological hypotheses without reducing the stringency of subsequent validation steps; those where it is used for replacing biological experiments with fictional data, yet with strong constraints on the generated data variance; and those targeting improved bioinformatic tools.
NB: This is the summary of an accepted paper (to be published in 2026), so I can go for an oral presentation, but as an SMPGD organizer I should step down for external contributions.

Keywords: omics data analysis, false discoveries, generative artificial intelligence, hallucination

*Speaker