glAIcobiology

Daniel BOJAR [1],

[1] University of Gothenburg, Sweden

daniel.bojar@gu.se

Computational research on glycans is frequently hampered by the sparsity, complexity, and diversity of glycan-related data. This issue is further compounded by the realization that, despite the dawn of Big Data, most glycan-related projects only yield rather modest quantities of data, compared to other systems biology disciplines.

We thus present a new paradigm of AI-aided glycoinformatics, with the ultimate aim of dispensing with the need for large datasets for functional inference. The key insight of, now widely used, models in other fields, such as AlphaFold or ESM, was that large, pre-trained models can be used to distil meaning from small datasets, or even single sequences in the case of AlphaFold.

Similarly, we argue that large, pre-trained models in glycobiology will substantially advance our understanding of glycan function. Concretely, we will present examples of such large AI models to elucidate, at scale: glycan measurement, glycan biosynthesis, and glycan function. All these efforts are fully open-access and facilitated by our glycoinformatics platform glycowork as well as numerous curated datasets for these purposes.

Overall, these advances clearly demonstrate that the field already contains sufficient data, albeit in a very scattered manner, to prepare large, pre-trained AI models that can then be used to accelerate progress in glycobiology.