Deep learning entering the post-protein structure prediction era : new horizons for structural biology

Sergei Grudinin

Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France;

sergei.grudinin@univ-grenoble-alpes.fr

The potential of deep learning has been recognized in structural bioinformatics for already some time, and became indisputable after the CASP13 (Critical Assessment of Structure Prediction) community-wide experiment in 2018. In CASP14, held in 2020, deep learning has boosted the field to unexpected levels reaching near-experimental accuracy. Its results demonstrate dramatic improvement in computing the three-dimensional structure of proteins from amino acid sequence, with many models rivalling experimental structures. This success comes from advances transferred from several machine-learning areas, including computer vision and natural language processing. At the same time, the community has developed methods specifically designed to deal with protein sequences and structures, and their representations. Novel emerging approaches include (i) geometric learning, i.e. learning on non-regular representations such as graphs, 3D Voronoi tessellations, and point clouds; (ii) pre-trained protein language models leveraging attention; (iii) equivariant architectures preserving the symmetry of 3D space; (iv) use of big data, e.g. large meta-genome databases; (v) combining protein representations; (vi) and finally truly end-to-end architectures, i.e. single differentiable models starting from a sequence and returning a 3D structure. These observations suggest that deep learning approaches will also be effective for a range of related structural biology applications that will be discussed in this lecture.

Keywords: deep learning, protein structure prediction, geometric learning, end-to-end architectures, protein language models