Oral presentation

Big box, small box, black box, George Box: some personal thoughts on determination and interpretation of structural models

Simon J. L. Billinge[1]

[1]*Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027*

sb2896@columbia.edu

A big discussion in the atomic pair distribution function (PDF) community seems to revolve around "big box" vs "small box" modeling. In this framing, a big box is a structural model that contains many thousands of atoms and typically uses a stochastic regression algorithm such as simulated annealing, and "small box" modeling has a greatly reduced number of atoms in the box, similar to crystallographic approaches of finding the unit cell and its contents. In these models, a local search is often carried out over a highly constrained set of parameters. There are clearly merits to both approaches. In this talk I will frame the question slightly differently in terms of what I call George Box modeling that takes inspiration from the writings of the famous 20th Century statistician, George Box. It gives us a slightly different way of looking at the question that may unlock some new insights. \n The session is also about "black boxes". I will take the discussion further to comment on black boxes, especially in the context of the recent push to apply machine learning (ML) approaches to crystallography problems. I will also reach out to George Box to help think about black boxes, but will also touch upon interpretable ML and how it can open small windows into black boxes, albeit slightly foggy and possibly distorted windows. I will also describe some of our efforts to apply ML to the nanostructure inverse problem: given a 1D measured PDF, what is the 3D arrangement of atoms that gave rise to it. Preliminary results seem to be very promising, for example the *DeepStruc* autoencoder deep neural network model that was developed together with the groups of Kirsten Jensen and Raghav Selven in Copenhagen [1] illustrated in the figure, but can the results be trusted?

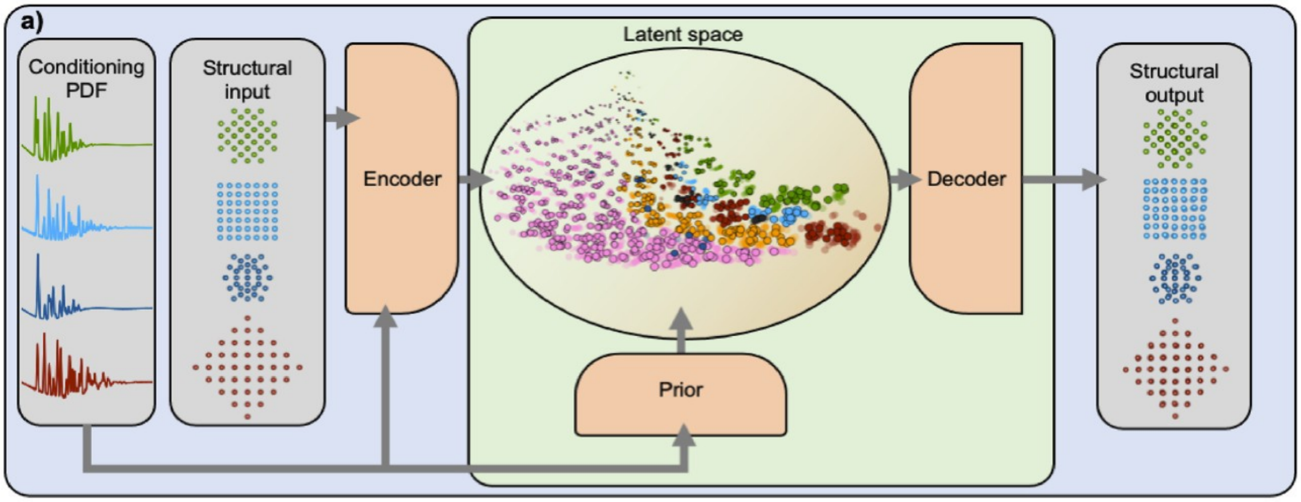

Figure 1. Schematic of a the *DeepStruc* deep neural net [1]. *DeepStruc* predicts the xyz-coordinates of input cluster structures with conditional inputs provided in the form of a PDF. The encoder uses the structure and its PDF as input while the prior only takes the PDF as input. To obtain the structural output a latent space embedding is given as input to the decoder which produces the corresponding output xyz-coordinates. During training of *DeepStruc* both the blue and green regions are used, while only the green region is used for structure prediction during the inference process.

[1] Emil T. S. Kjær, Andy S. Anker, Marcus N. Weng, Simon J. L. Billinge, Raghavendra Selvan and Kirsten M. Ø. Jensen. “DeepStruc: Towards structure solution from pair distribution function data using deep generative models”. In: Digital Discovery 2 (2023), pp. 69–80. doi: 10.1039/D2DD00086E