UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

Qualifying Exam

Monday April 8, 2024 at 1PM, online via Zoom (Remote)

Angela Kohut

[3]

Using Encoded Protein Structure Representations and Transfer Learning for Protein Structure Applications

Chair: Dr. Stacey Scott
Advisor: Dr. Stefan Kremer
Co-Advisor: Dr. Steffen Graether (Biophysics)
Non-advisory: Dr. Yan Yan
Non-Advisory: Dr. Leonid Brown (Physics)

Abstract:

Proteins play a crucial role in various biological processes, serving as the building blocks of life. Therefore, understanding their structure and function is paramount for advancing our knowledge of structural biology. The Protein Data Bank (PDB) files have been an integral part of helping researchers decipher the complex workings of proteins. PDB files provide three-dimensional Cartesian coordinates of protein structures used as a stepping stone for other protein structure tools, such as protein classification.

Efficient protein classification is vital for organizing and categorizing the large number of proteins discovered to date. It enables researchers to identify functional relationships, predict protein functions, and gain insights into their evolutionary history. However, current protein structural classification systems like CATH and SCOP have some drawbacks, such as the need for manual curation in some or all classification steps, challenges in handling the increasing number of protein structures, and the rigidity of classification definitions.

Recently, image processing has advanced significantly, mainly due to neural networks. Deep learning networks, such as Convolutional Neural Networks (CNNs), consist of multiple layers, each capturing distinct information. Layers near the input reflect provided data, while those near the output hold task-specific training details. Transfer learning is vital in vision learning systems as it leverages insights from initial layers, which transfer to other tasks or fields. CNNs have transformed tasks like image recognition and found successful applications in various areas like computer vision and understanding natural language.

This work combines image processing and transfer learning to develop new encodings from three-dimensional protein information for various classification systems. Given the remarkable success of learned representations and CNNs for image processing, I am confident that this approach can develop representations useful for classification. The protein encodings will be helpful in other protein structure-related problems such as protein structure prediction, protein function prediction, and drug discovery.