# 

# Cyclica AlphaFold 2 Challenge

A UWaterloo Data Science Club data hackathon challenge hosted by Cyclica.

## Objective

Predicting drug binding sites on AlphaFold2-predicted proteins.

## Introduction

[AlphaFold2](https://www.deepmind.com/research/highlighted-research/alphafold) (AF2) is arguably one of the biggest scientific breakthroughs of the past decade, offering accurate AI-based 3d structure predictions for [all human proteins](https://alphafold.ebi.ac.uk/download), as well as 47 other model organisms and global health species. However, there remains a long road ahead in translating these predictions into new medicines. First, pharmaceutical drug discovery applications require an understanding of where a small molecule drug can physically bind to the surface of a 3D protein structure. In this challenge, we will be predicting probable drug-binding sites for thousands of human proteins, thereby mapping the landscape for possible medicines!

## Background Knowledge

[Proteins](https://en.wikipedia.org/wiki/Protein) are the product and physical manifestation of genes in all living organisms. They have formed as long chains, where each link is 1 of 20 [amino acids](https://en.wikipedia.org/wiki/Amino_acid) (aka residues). In the cell, these protein chains [fold](https://en.wikipedia.org/wiki/Protein_folding) into specific 3D shapes, forming all kinds of molecular machines that drive life through an incredible assortment of biological tasks, including tiny molecular motors, generators, pumps, glues, transporters, reactors, containers, propellers, and many, many, many more!

## Dataset

Cyclica has mapped all verifiable drug binding sites onto AlphaFold2 protein structures. We then labeled each protein residue as 'drug binding' or 'non-drug binding' accordingly. Moreover, we have generated data-rich features for each residue in the dataset, derived exclusively from the AF2-predicted protein structure using standard structural biology tools. See cells below for further information.

## Challenge

1. Develop a classification model to predict 'drug binding' or 'non-drug binding' for any query residue.
2. Defend the model by demonstrating which features are most important for drug-binding.

### **Submission**

* As with all other challenges, ensure that you have submitted the Google form, the Devpost
* An sample CSV to be submitted to the Google Form can look like this:

| ,Predicted 0,0 1,0 2,1 3,0 4,1 5,0 ... |
| --- |

*Note: specific knowledge of protein structure or biochemistry may be necessary to participate. The accompanying dataset should have all properties available to generate predictive models.*

## Prize

* $500 Gift Card.
* A promoted feature on Cyclica's blog, widely seen in the biotechnology and AI fields.

## About Cyclica

Cyclica is a neobiotech company fueled by data-driven drug discovery. We advance molecules that embrace the complexity of disease. Our work spans dozens of collaborations with large pharma and biotech as well as 50 joint ventures and counting. We are a passionate team of biotech and pharma professionals, biologists, chemists, and computer scientists who live and labor at the intersection of our collective expertise. For more information about our company, please visit: [www.cyclicarx.com](https://colab.research.google.com/drive/www.cyclicarx.com).

## Additional Resources

* [EMBL-EBI AF2 Downloads](https://alphafold.ebi.ac.uk/download): Download 3D structures for all human proteins, or all proteins from a collection of 48 different model organisms and global health proteomes.
* [PDB Molecule of the Month](https://pdb101.rcsb.org/motm/motm-by-date): Learn about new proteins and how their 3D structures impact function.
* Cyclica's [Blog](https://blog.cyclicarx.com/) and [LinkedIn](https://www.linkedin.com/company/cyclica/): See content on related to our NeoBiotech model and science, including our AlphaFold2 series.