Week 3 (page 1)

Lecture 6

Open science bioinformatics, binding site prediction with NNs

Session 6.1

OPEN SCIENCE IMPORTANCE FOR STRUCTURE PREDICTION

Importance of Open Science wealth of structural data and genome, proteome information; Development of AlphaFold.

Session 6.2

DEEP SITE BINDING SITE PREDICTION

Introduction to DeepSite, Binding Site Prediction using Neural Network, application and method. Considerations in ligand macro-molecule binding-site prediction.

Quiz 6

UNDERSTANDING QUESTIONS

(6 for each session) to answer at Studium.

Exercise 6

AI FOR STRUCTURE AND BINDING SITE PREDICTION COMPARED WITH KNOWN STRUCTURE FOR PHOSHOLIPASE A2

Method: 3D Viewing, Convolutional Neural Networks, DeepSite
(Narrative here. The exercise itself is on Studium)

❒ Graded material

✹ Extra understanding

MACHINE LEARNING IN STRUCTURE PREDICTION, A TECHNOLOGY ENABLED BY THE OPEN SCIENCE COMMUNITY

FIGURE: Logos from a few example contributors to open science and ML structure prediction.

AlphaFold was developed by the Google company Deepmind together with the EMBL-EBI (European Molecular Biology Laboratory – European Biological Institute) and other organisations such as PDB, Uniprot etc and peers who have made the current progress possible by the open-science community.

AlphaFold protein structure predictions are covering the proteomes of at least 21 species and growing, and have been made freely available to the scientific community, and they are also accessible through the Uniprot website through a structure viewer.

Open data resources like PDB and UniProt have been key in the successful development of AlphaFold. AlphaFold structure predictions are now freely accessible to the scientific community and is a success story for academia-industry collaboration and for open science. The Universal Protein Resource (UniProt) is a comprehensive resource for proteome related data, where one can access much bioinformatical data (The Uniprot Consortium, 2002-2022).

THE INCREASING STRUCTURES THROUGH THE YEARS

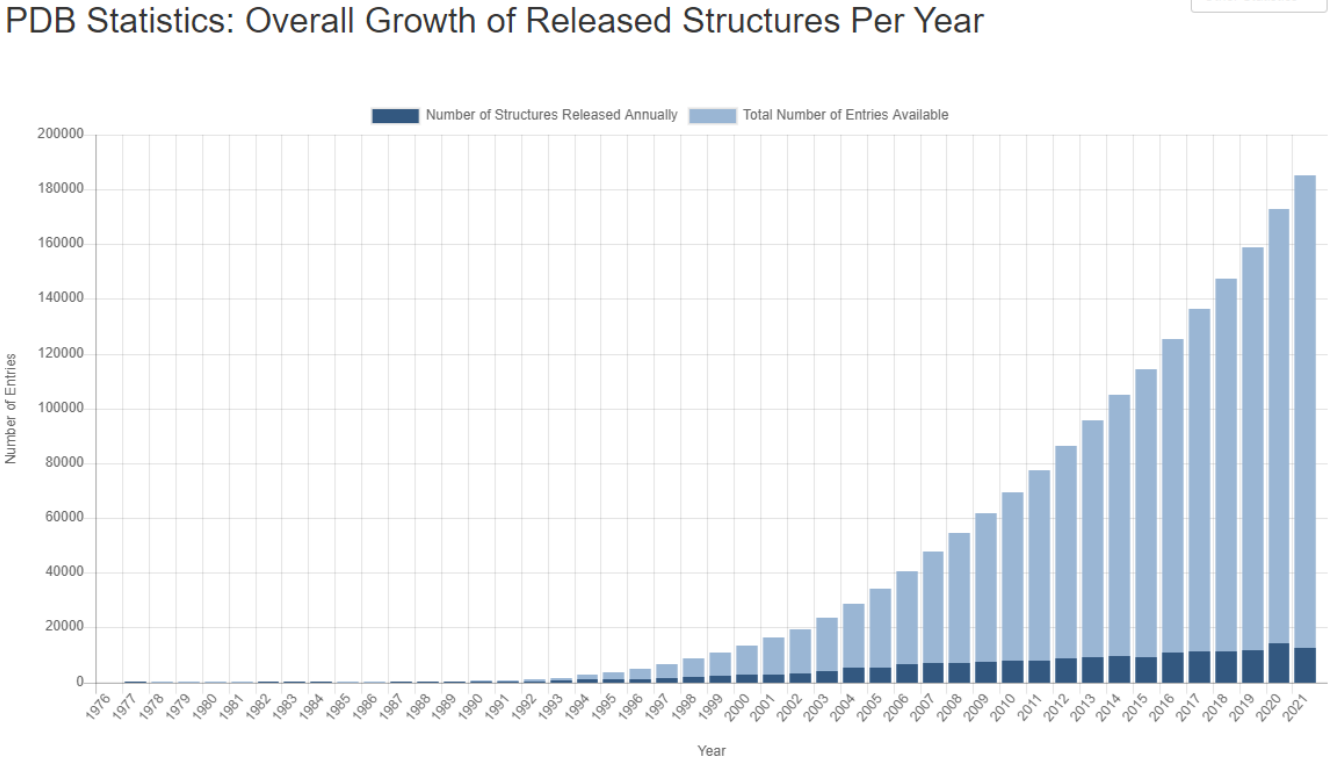


FIGURE: A diagram of the number of published structures in Protein Data Bank per year between year 1976 to 2021, (PDB Statistics: Overall Growth of Released Structures Per Year, no date).

Over the past 50 years, we've seen an increase and growth of total number of structures available (light-blue), and each year the number of published structures (dark blue) has also increased.

Over the next 50 years, we can expect the PDB archive to grow very much in size and importance as it provides ever larger and more diverse communities of data consumers with open access to well-validated, expertly biocurated 3D structures of biological macromolecules.
Protein data bank is the main repository for structural biology data (Berman, 2000), yet many other data-bases exist that are in collaboration and utilize the same structures.

Many experimental methods have been used to curate structures. Most structures are obtained with X-ray crystallography, NMR, and in later years with Cryo-EM techniques. Novel techniques are constantly in development.

Many years of painstaking and laborious work and multi-million dollar financial investment over the years have been done to solve the structures that are currently in the PDB data-base.

AVAILABILITY OF PDB TRANSFORMS UNDERSTANDING

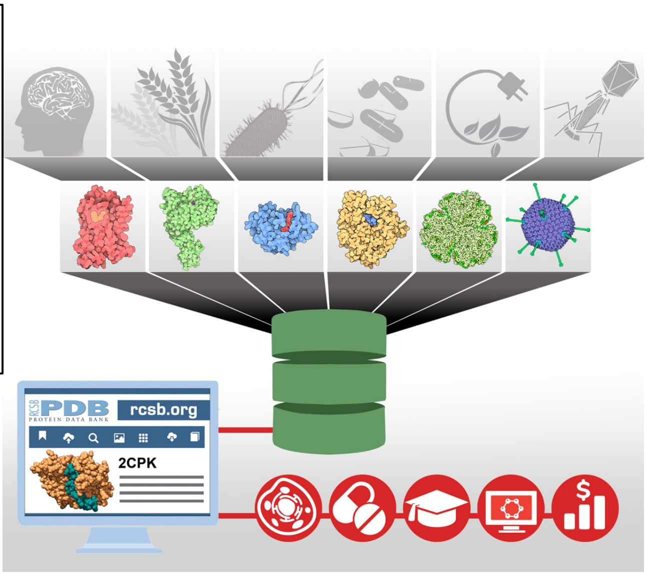


FIGURE: Schematic of the importance of the PDB to various relationships have to different fields of research and understanding of life-processes, not the least for drug design. Image from (About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education, n.d.).

In the early 1980's, it was for many structural biologists an unrealized goal and lack of ability to rationally design drugs using protein structures. The first project were underway in the mid-80's, and by the early 1990s the first successes were published. (Anderson, 2003)

Open access to Protein Data Bank (PDB) archive, was established in 1971, and has for decades benefited basic and applied research and education, and has helped transform 3D structure data to knowledge.

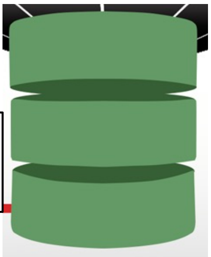
Over the past decades, structural biology and structure-guided drug discovery, have become firmly established within the biopharmaceutical industry since the field can explain how small-molecule ligands bind to their target proteins. The field has also proven useful in solving some of the countless challenges that is inherent in turning biochemically active compounds into potent drug-like molecules that is suitable for safety and efficacy testing in animals and humans. (Burley, 2021)

The field of structural-based drug design, has been growing, and many successes have occurred in recent decades and in recent years. It is expected a growing trend also in how that structural data and ML will aid our understanding of structural-based drug design and what the impacts of drugs have on human health.

The increase in structural information and other biological information, have provided hundreds if not thousands new targets and opportunities for future drug lead discovery.

OPEN ACCESS DATA: A CORNERSTONE FOR AI APPROACHES TO PROTEIN STRUCTURE PREDICTION

FIGURE *1*: Open source and standardized data is a corner-stone in ML prediction.



The Protein Data Bank was established in 1971 to archive 3D structures of biological macromolecules for the good of society. Even fifty years later, the PDB are serving millions of data consumers around the world with open access to more than 175,000 experimentally determined structures of macro-molecules and their complexes with one another and small-molecule ligands.

The vast wealth of 3D structure data stored in the PDB has resulted in significant advances in our understanding of protein architecture, culminating in breakthroughs in protein structure prediction accelerated by artificial intelligence approaches and machine learning methods.

In aftermath, we can appreciate the importance of having all open access to structural biology data expertly validated, curated and made available from a repository with a standardized format. The progress in structural biology would have been agonizingly slow in a world without open-access to data. We would then not have seen as many 3D structures becoming key drivers of research progress across the sciences and in structure-based drug discovery.

In contrast, PDB data pertaining to small-molecule interactions with proteins has been limited by data being behind the firewalls of biopharmaceutical companies. The deposition of more data to PDB from industry would almost certainly fuel advances in prediction of small-molecule binding to proteins and nucleic acids. We can expect acceleration of drug discovery and development efforts in both academia and industry, with more sufficient data in the public domain.

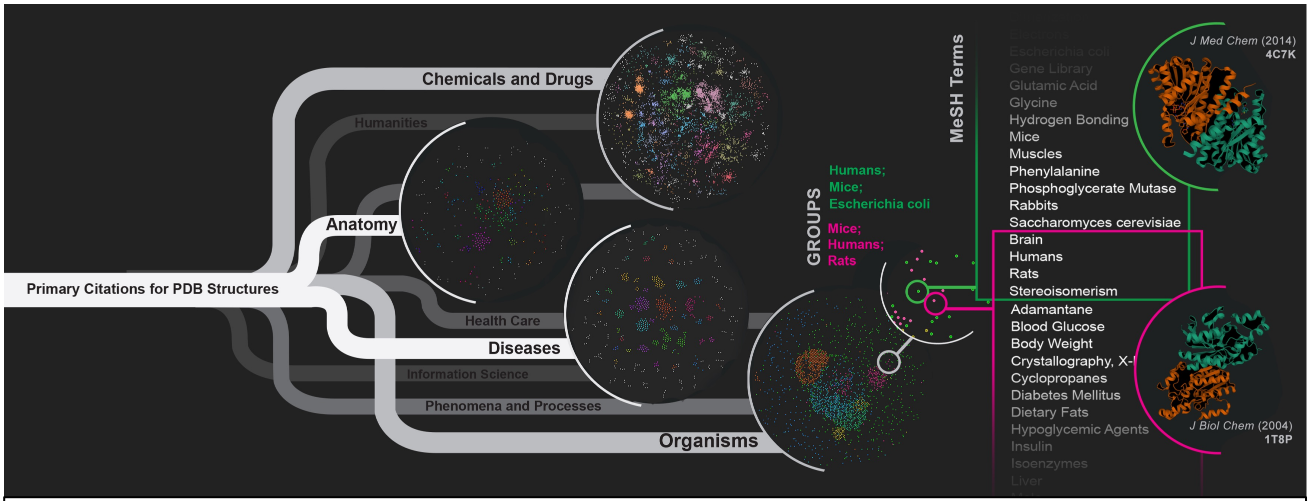


FIGURE 2: Overview schematic of PDB Citation MeSH Network Explorer, finding connections between articles describing PDB entries (RCSB, no date).
The year 2021 marks the 50th year of the PDB, as of january 2021, the PDB contained more than 170,000 structures; ~150,000 had corresponding “primary citations” describing these entries in a peer-reviewed journal.
The National Library of Medicine assigns MeSH (Medical Subject Headings) from a controlled vocabulary to index articles for PubMed. MeSH terms typically appear in a hierarchical tree structure that starts with 16 main branches (lettering A-P), e.g. A. Anatomy, B. Organisms, C. Diseases, D. Chemicals and Drugs etc. RCSB PDB has a separate browser to find PDB structures based on these hierarchical trees at RCSB.org. The ”RCSB PDB Citation MeSH Network Explorer” flattens these trees into co-occurence networks of MeSH terms associated with PDB entries. Each node on the graph is a publication, and nodes are linked when they share MeSH terms.

In late year 2020, DeepMind revealed that its Alphafold2 system can predict 3D structures of small globular proteins with accuracies comparable to those of low-resolution experimental methods. Given the success of the earlier Alphafold version, it was not unexpected that the performance of Alphafold2 to be heralded as a major breakthrough in de novo protein structure prediction.

As estimated in year 2021, it is uncertain the degree to which ML methods will be able to reduce prediction errors and expand its horizon to larger, multidomain proteins. We can, however, assert confidently that the open-source sharing of methodologies and code by DeepMind and other organisations using ML approaches, will be vital to the long-term success of these endeavours.

Even if AI methods themselves would not improve, year-upon-year growth of the PDB is likely to improve prediction accuracy for small globular proteins. Likely as the volume and standardisation of open access structure data continues to increase, accuracy will improve and the impact of AI advances will broaden. Text *content from Burley and Berman (2021).*

ALPHAFOLD PREDICTS STRUCTURES OF WHOLE PROTEOMES

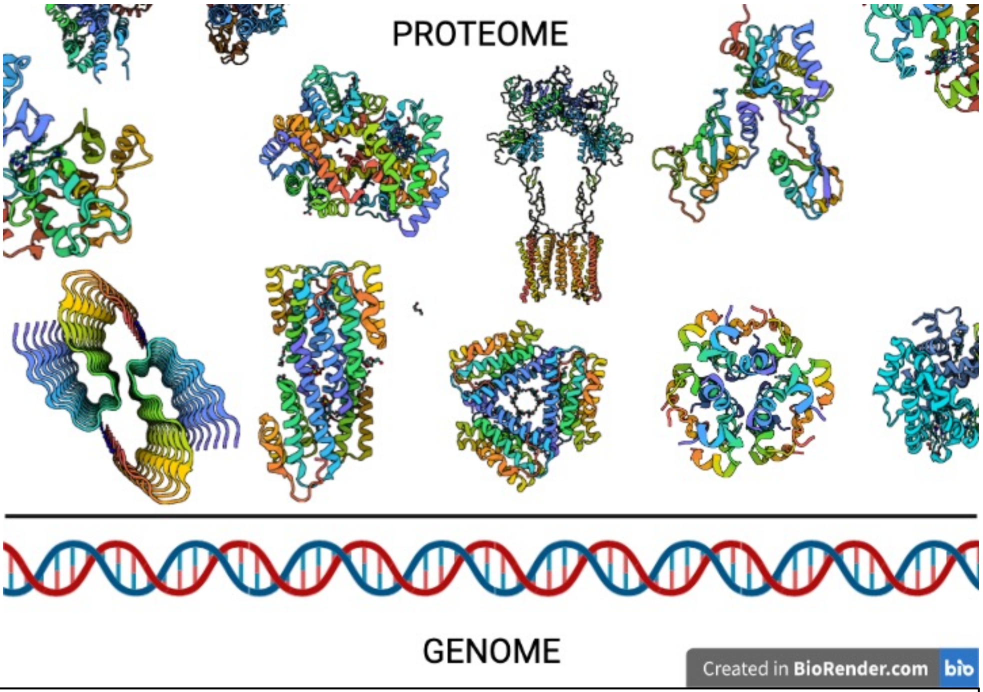


FIGURE : Depiction of the protome (the whole collection of proteins) and the genome (the whole collection of genes coding for particular proteins).

AlphaFold markedly expanded the proteome structural coverage by applying state-of-the-art ML method, AlphaFold2, at a scale that covers 98.5% of human proteins (almost the entire human proteome of 20,000 proteins). Predicted structures for the UniProt human reference proteome (with one representative sequence per gene), however with a number of amino-acid residues upper length limit. (Tunyasuvunakool et al., 2021).

A partnership with DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) created the AlphaFold DataBase to make freely available these predictions to the scientific community. The first release covered the human proteome and the proteomes of several other organisms key for studying, while the second release added the majority of manually curated UniProt entries. Planned to in year 2022 and beyond, extend the prediction to a large proportion of all catalogued proteins, over 100 million. (DeepMind and EMBL-EBI, 2021).

LOCATING AREAS WHERE A DRUG SHOULD BIND

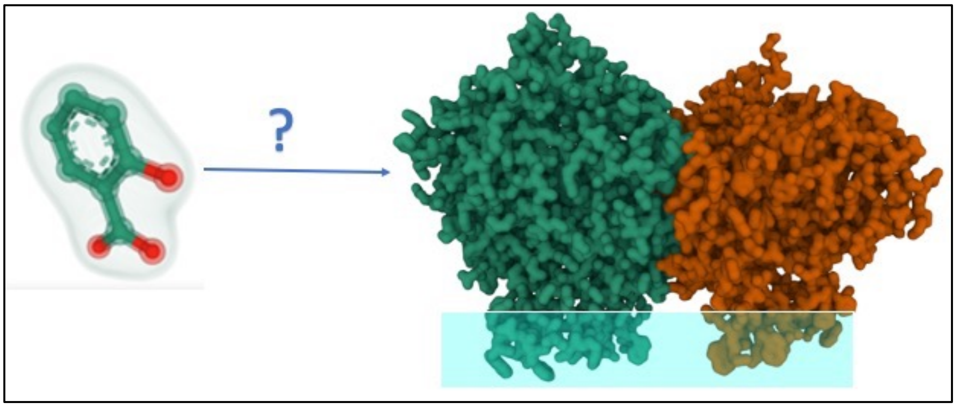


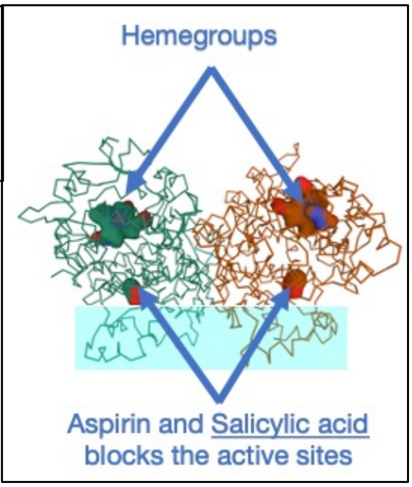
FIGURE: To the left is an representation of aspirin, the carbons are arbitrarily coloured in green and the oxygens in red. To the right is an all-atom structure of a enzyme cyclooxygenase (from PDB 1PTH), where all atoms in the monomers in the dimer are arbitrarily coloured in green and orange respectively. The two protein monomers in the dimere are combined making up the protein cyclooxygenase. There are identical binding pockets inside each of the two monomers.

The choice of a drug target, is primarily made on a biological and biochemical basis. The ideal target macromolecule for structure-based drug design is one that is closely linked to human disease and binds a small molecule in order to carry out a function. The target macro-molecule usually has a well-defined binding pocket.

Structure-based drug design begins with the identification of a potential ligand binding site on the target macro-molecule. Ideally the target is a pocket or protuberance with a variety of potential hydrogen bonds donors and acceptors, hydrophobic characteristics, and sizes of molecular surfaces. The ligand binding site can be the active site, as in an enzyme, an assembly site with another macromolecule, or a communication site necessary in the mechanism of the molecule. Note that, in addition to the well-accepted protein target sites, RNA secondary structural elements can provide excellent target sites since they are species specific, bind ligands, and can be specific for a disease state. *(Anderson, 2003)*.

Among the most important steps in structure-based drug design is identifying viable druggable binding sites on the target protein. This step is defined as identifying and discerning in the protein hollow areas, potentially at the surface, however possibly in hidden surfaces, which may be likely to bind to a small compound.

FIGURE: Ribbon structure of enzyme cyclooxygenase (from PDB 1PTH), showing where the heme-groups are located and also where the ligand aspirin and salicylic acid binds and blocks the active site.



Atleast two factors can be considered to define a ”druggable” protein target:

- One factor is concerning an important aspect: the ability of a protein target macro-molecule to bind a small drug-like molecule.
- Second factor is whether its binding has a desired therapeutical and health effect. Which will also depend on the analysis of the cellular pathways requiring a systems biology understanding.

For larger macro-molecular proteins it can be more difficult to solely locate or discern its mechanism from structure and predicted binding pockets alone. Cyclooxygenase has two particular active sites known from empirical experiment and from knowledge about its structure, that aspirin has been inferred to inhibit, these are somewhat hidden within the macro-molecule. Whereas the heme-groups also are bound ligands with a particular adjoining function in the cyclooxygenase reaction.

All structure of biomolecules and target proteins such as cyclooxygenase has a function and some parts have known functions, whereas other parts of the protein may still have unknown or yet to be clarified functions in the biochemical machinery. It can hence be difficult to be aware of all function or binding areas or all the biological molecules or ligands that can affect the dynamics of the macro-molecule or all the cellular pathways affected. Literature is necessary to compare with.

Monomer: a molecule of any of a class of compounds, mostly organic, that can react with other molecules to form very large molecules, or polymers. When referring to protein monomers, one means that several protein macro-molecules can assemble and form a quaternary structure, for example a dimer is then a protein that consists of two identical protein monomers.

DEEPSITE BINDING POCKET PREDICTION USING 3D-CNN

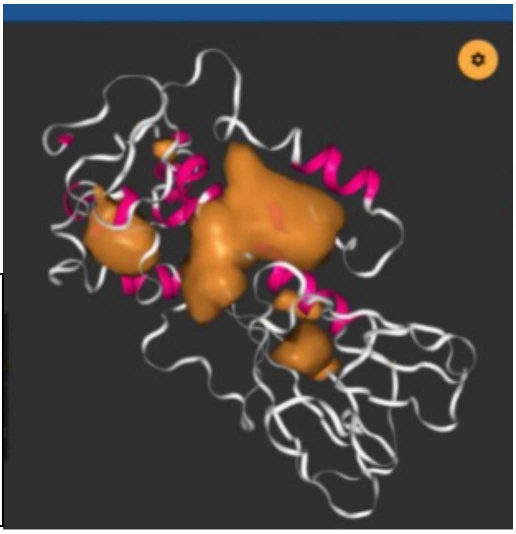


FIGURE: DeepSite, a binding pocket predictor using neural-networks, predicting drug binding sites on protein macro-molecules, using a three dimensional convolutional neural network (CNN) (Acellera, 2017).

It has been determined experimentally that aspirin and salicylic acid binds to two identical binding pockets inside the cyclooxygenase monomers (Loll et al. 1995). Is there a way to predict plausible binding sites or pockets for molecules from empirical data with neural networks?

Well, for detecting binding cavities, those likely to bind to a small drug compound, several algorithms has been developed over the years by considering cleverly the geometric, chemical and homolog structure features of proteins. Prediction of druggable binding sites is an important step in structure-based drug design.

DeepSite, is a pure ML algorithm for predicting protein ligand binding sites learning purely from examples and without encoding any problem specific knowledge. The algorithm has been extensively tested according to different criteria, with the conclusion that it was able to outperform other existing state of the art strategies.

DeepSite was trained on the scPDB database, which is a comprehensive and updated selection of ligandable binding sites of structures in the PDB. The binding sites are defined from complexes between a pharmacological ligand and a protein. The scPDB database provides the all atom description of the ligand, the protein, and the ligand binding site at protein, and the binding mode. *Text from Jimenez et al. (2017).*

Homolog structure features of protein: When similar macro-molecules appear in different species of organisms, they are called homologous, for instance the protein insulin have similar structure but slight variation in amino-acid sequence and structure in various species.

THE DEEPSITE ALGORITHM

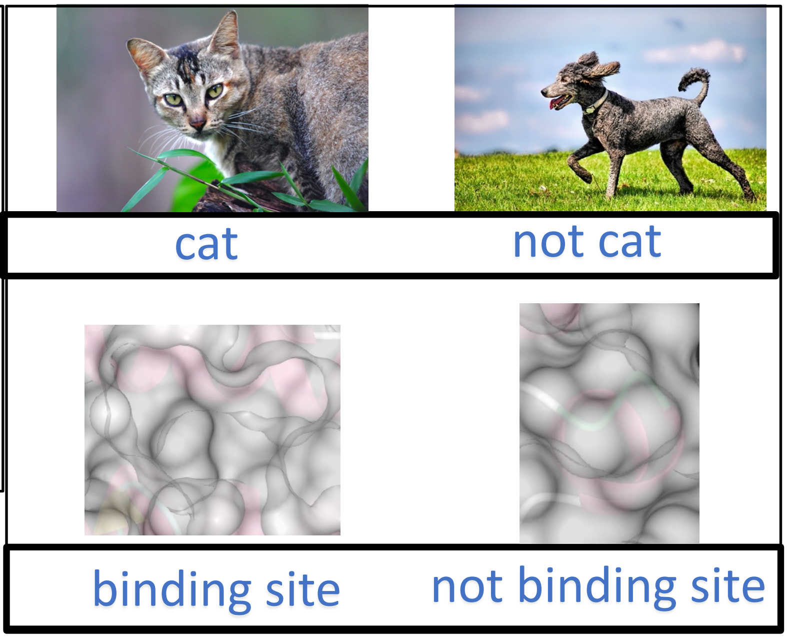


FIGURE 1: Illustration of DeepSite pocket and not-pocket detection analogy with how 2D-CNN can distinguish cat from non-cat. Cat & dog pictures: from Creative Commons. Green, blue, red version of cat, made with GIMP-2.10. Binding site & not binding site: Rendered in DeepSite from structure 1BP2 (phospholipase A2).

DeepSite leverages from the development of deep learning in the image recognition field and brings it to structural biology.

For deep learning to function in classifying 2D images, neural networks need to be trained with multiple data examples of something we want to learn. In a classification problem, we may want the NN algorithm to be able to discern and correctly differentiate new examples of the classification 1 (say cats) from the classification 2 (say not-cats).

The algorithm of DeepSite is not so different from a simple image 2D classifier. However it treats 3D-images of atomic structures of ligand and macro-molecule complex. What DeepSite does is that it takes examples of binding sites and non binding sites from the scPDB database and trains a neural network (NN) to correctly classify what looks like a ligand binding site from what does not look like one.

Since the structures of proteins are three-dimensional and to use as much information as possible, it is not used simple 2D pictures, actually it is used a 3D box that contains a fraction of the protein structure. In practice, this means it is used a type of neural networks called 3D convolutional neural networks, which are different from the 2D convolutional networks used for image recognition, since images are two-dimensional.

Moreover, in 2D image recognition, the pixels having information about color is decoded and input into the NN in the form of 3 channels (red, green, blue) that combined can reconstruct each pixel of the original image. For instance, this cat picture consisting of a 2D array of pixels can be classified into the three channels of colour, and also reconstructed based upon these channels.

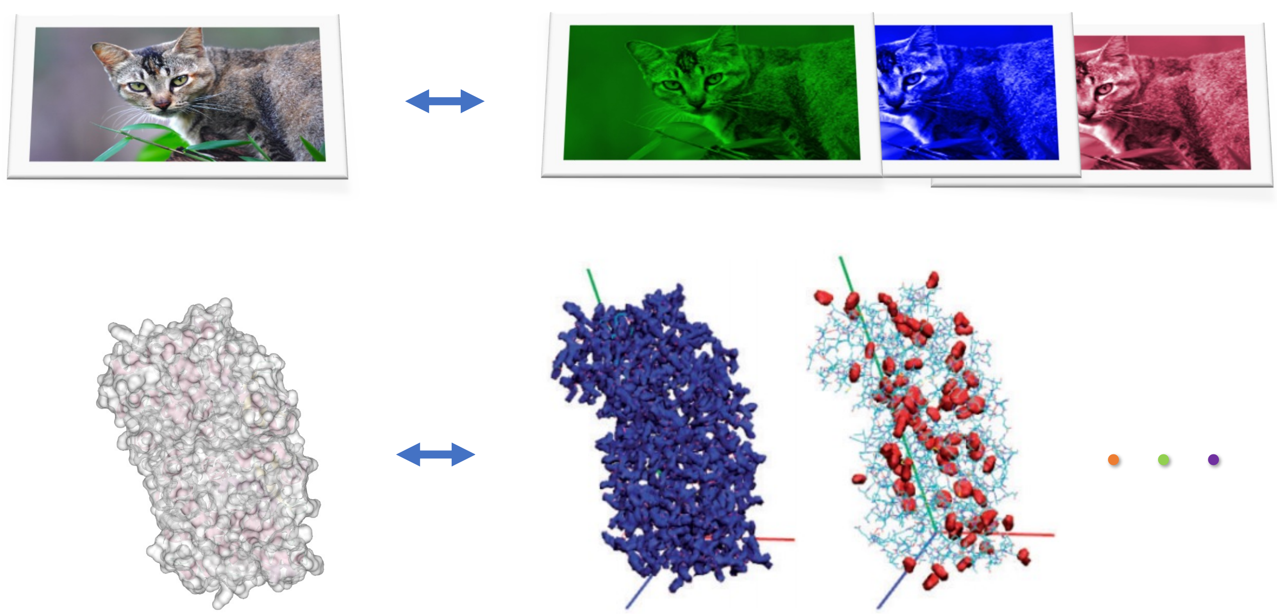


FIGURE 2: Top, the cat image separated into 3 channels. Bottom, the PDB protein structure 4NIE classified into 2 example channels of atomic chemical properties, hydrophobicity and aromaticity. PDB protein structure 4NIE rendered in DeepSite and hydrophobic/aromatic picture are from referenced article (Jimenez, 2017).

In 3D image recognition, the equivalent of a pixel in 3D is used called a voxel, here the same principle is used to define the information-content of any boxes containing atomic structures making up "binding sites" to "not binding sites", where the atoms of each boxes are classified into several channels, each channel accounting for atoms that have different chemical properties: Hydrophobicity, Aromaticity, H-bond acceptor, H-bond donor, Positive ionizable, Negative Ionisible, Metal, Excluded volume.

This information of channels with atomic structure to chemical property association was generated for 7622 ligand bounded proteins from the scPDB database for examples of boxes containing atomic structures classifying boxes as "probable binding site" and "non probable binding site" and a 3D-CNN is trained to differentiate these two classes. Content references (Jimenez et al., 2017), (Gerard Martinez, 2017).

Pixel: a word invented from "picture element" being a basic unit of programmable color on a computer display or in a computer image. The size of a pixel depends on how the resolution is set.

Voxel: In 3D computer graphics, a voxel represents a value on a regular grid in three-dimensional space. As with pixels in a 2D image, the size of the voxel depends on the resolution set.

USING THE TRAINED DEEPSITE 3DCNN MODEL

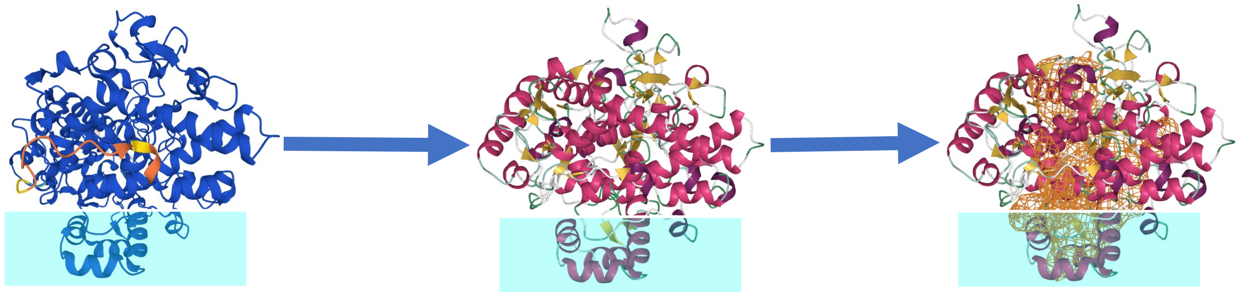


FIGURE: Illustration of testing the trained 3DCNN model,
Binding site prediction from DeepSite of one of two identical monomers in a dimer of the protein enzyme cyclooxygenase (Uniprot ID P23219) an larger macro-molecule. Light-blue just drawed upon illustratively for the part otherwise embedded in cell-membrane.
Left: AlphaFold prediction cyclooxygenase (Uniprot ID P23219).
Middle: Uploading this structure to DeepSite at playmolecule.
Right: The resulting predicted binding sites or volumes that DeepSite predicts.

Thereto, using the trained 3D-CNN model, is when one submit a protein to the DeepSite web-application at playmolecule, wherein the model divides the protein atomic structure into boxes and for each box it asks the pre-trained model: is this a "probable binding site"? From which the 3D-CNN model gives out a probability of being a "binding site": that ranges from 0 (likely "not binding site") to 1 (likely a "binding site"), e.g. 0.5 would mean approximately "50% probability of being a binding site".

The calculated "probability in 3D space" is aggregated into a so called isosurface, which is an interpolation of the different probabilities obtained from the predictions of each box of atomic structure. The isosurface which is on the rightmost structure (shown in orange grid-wires), helps to highlight which area is more likely to be a binding site, and it is built in to the toolset of the DeepSite web-viewer to vary the iso-surface shown. Text references (Gerard Martinez, 2017).

Cyclooxygenase is hollow and DeepSite predicts many areas of binding, including the regions where the ligand heme group is bound, in addition ot the pocket where aspirin binds, and also the tunnel from whence the active site is reached from the membrane side where this protein is embedded in.

Initially, this structure is a prediction from AlphaFold which has high confidence in most of its parts. Different structures of the same protein used in binding pocket prediction may produce different pictures based also upon how much the structure reflects its biological reality in the cell, by the same token depends on the database structures used for training the ML model. In any case predicted binding sites has to be compared with systems biology, experiment and empirical knowledge about the macro-molecule, ligand and affected cellular pathways.

Understanding question to answer at Studium

Session 6.1

6 UQ □

Session 6.2

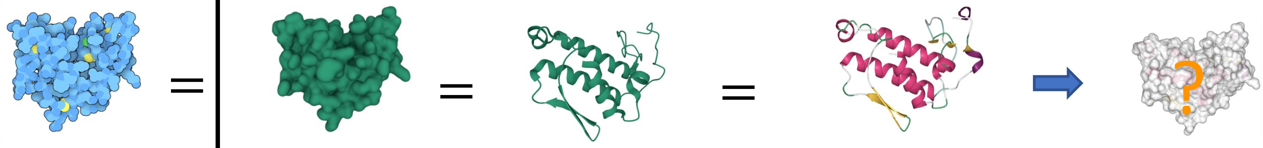
6 UQ □

AI FOR STRUCTURE AND BINDING SITE PREDICTION COMPARED WITH KNOWN STRUCTURE FOR PHOSHOLIPASE A2

METHOD: 3D VIEWING, CONVOLUTIONAL NEURAL NETWORKS, DEEPSITE

(Narrative here. The exercise itself is on Studium)

Question 1 & 2



NARRATIVE

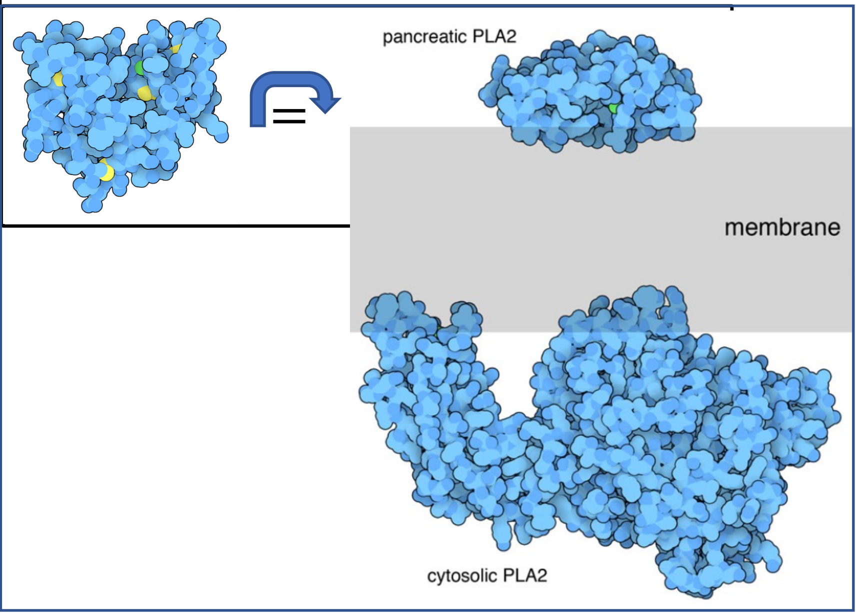
FIGURE**:** Graphical abstract *of exercise. All depictions of this protein shown with the same orientation and view-angle.*

- ***NARRATIVE INFORMATION***
  First we present some overall knowledge of this drug-target macromolecule protein phospholipase A2.

- ***NARRATIVE Question 1 & 2***
  Here we will investigate the 3D structure using the inherent 3D viewer at the protein data bank. We will investigate the macro-molecule, then let this structure go through the binding site prediction of the DeepSite 3D-CNN model.

EXERCISE 6 - **NARRATIVE INFORMATION**

FIGURE: Illustration of phosoholipases. Extra-cellular protein phospholipase A2 in black frame, which we will study in the exercise, a bound calcium ion in the structure is shown in green. Moreover, a larger cytosolic variant visualized below membrane. Image from Goodsell (2019).



Some types of phospholipases are secreted, such as the phospholipases made by the pancreas for use in digestion. Others are made inside cells, where they help with creation of signaling molecules. All of them share a similar function of chopping off one of the tails of phospholipids.

Many various forms of phospholipase A2 are made for different functions in the body. These enzymes all have areas on their surface that interact with the face of a membrane, allowing the enzymes to extract lipids for cleavage. Secreted phospholipase A2 is small in size. In contrast, the cytoplasmic enzymes are larger (shown in IMAGE below "membrane", having separate domains that are involved with interaction with membranes and catalyzing the cleavage reaction (shown here from PDB entry 1CJY).

Phospholipases are small enzymes that can withstand the harsh environment when they are secreted (or injected) outside of cells. They are tied together by many bonds between amino-acids called cysteines, these bonds stabilize the folded structure. The one shown here within black frame, from cow pancreas (PDB entry 1BP2), and the other one in following page human PLA2 (Uniprot P04054), have seven bonds between cysteines. The active site is a pocket at one side of the protein, with a calcium ion that catalyzes the cleavage reaction of phosholipids.

Some of the lipid tails that are released by phospholipase A2 are used to build signaling molecules involved in inflammation and pain. Because of this action, disorders in phospholipase action can contribute to diseases such as atherosclerosis and Crohn's disease. The community of researchers are currently using structures of phospholipases to discover new drugs to cure these diseases by blocking the action of the enzyme. Text from Goodsell (2019).

EXERCISE 6 - **NARRATIVE Question 1 & 2**

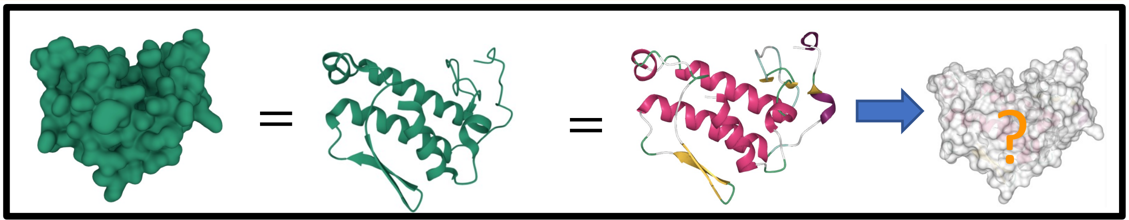


FIGURE: Graphical abstract Question 1 of Exercise 6
RCSB PDB (1BP2) structure of phospholipase A2, from bovine (cow) origin, same orientation and view-angles.

One can access this structure at https://www.rcsb.org/3d-view/1BP2. One can play around with the graphical user interface at the Polymer and Ligand settings. **There is a strongly bound calcium ion here, which may be indicative of an active site of this enzyme.**

The first structure from the left in green is an atom-atom (Gaussian Sphere) surface view of the protein which is a folded chain of 123 amino-acids.

The second structure from the left in green is a ribbon representation of the backbone of this 123 amino-acid protein, hence no atoms shown.

The third structure from the left in purple/yellow/white is just the protein structure uploaded to the web-application DeepSite, note however even if only the ribbon backbone representation is shown, DeepSite considers all atoms when predicting binding pockets.

The rightmost structure is a depiction from DeepSite prediction of this structure

The exercise will focus on predicting and affirming the binding site, where is it ?

Please follow the instructions at Exercise 6 at Studium to complete the Questions

References (Harvard citation style)

About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education, n.d.. Available at: https://www.rcsb.org/pages/about-us/index (Accessed: 31 Mars 2022).

Acellera, 2017. DeepSite: a binding pocket predictor using neural-networks, Available at: https://www.playmolecule.com/deepsite/ (Accessed: 1 April 2022)

Anderson, A.C., 2003. The process of structurebased drug design. Chemistry & biology, 10(9), pp.787-797.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. *Nucleic acids research*, *28*(1), pp.235-242. https://doi.org/10.1093/nar/28.1.235

Burley, S.K., 2021a. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. *Journal of Biological Chemistry*, 296. https://doi.org/10.1016/j.jbc.2021.100559

Burley, S.K. and Berman, H.M., 2021. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. Structure. https://doi.org/10.1016/j.str.2021.04.010

Clementi, C., 2021. Fast track to structural biology. *Nature Chemistry*, *13*(11), pp.1032-1034. https://doi.org/10.1038/s41557-021-00814-y

DeepMind and EMBL-EBI, 2021. AlphaFold Protein Structure Database, Available at: https://alphafold.ebi.ac.uk/ (Accessed: 31 Mars 2022

Goodsell, D.S., 2019. Phospholipase A2, RCSB PDB Molecule of the Month by David S. Goodsell. Available at: https://pdb101.rcsb.org/motm/239 (Accessed: 31 Mars 2022).

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A.S. and De Fabritiis, G., 2017. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, *33*(19), pp.3036-3042. https://doi.org/10.1093/bioinformatics/btx350

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589. https://doi.org/10.1038/s41586-021-03828-1

Martinez G., 2017. Searching for binding pockets using PlayMolecule® DeepSite [TUTORIAL]. Available at: https://medium.com/playmolecule/searching-for-binding-pockets-using-playmolecule-deepsite-tutorial-dc08316eb464 (Accessed: 31 Mars 2022)

PDB Statistics: Overall Growth of Released Structures Per Year, n.d. . Available at: https://www.rcsb.org/stats/growth/growth-released-structures (Accessed: 31 Mars 2022).

RSCB, no date, PDB Citation MeSH Network Explorer, Available at: https://cdn.rcsb.org/rcsb-pdb/pdb50mesh/index.html (Accessed: 31 Mars 2022)

The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, https://doi.org/10.1093/nar/gkaa1100

The Uniprot Consortium, (2002-2022). Available at: https://www.uniprot.org/ (Accessed: 31 Mars 2022)