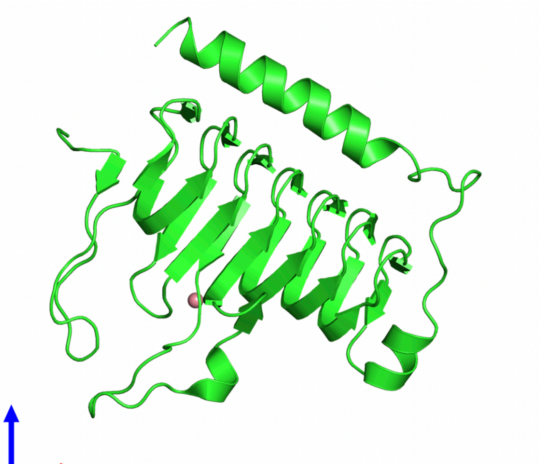
Notes for Maverick

Enzyme Sequence

MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNVQDGVVLHALETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSQVHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTNEAVVYVNVHLAEGYKETS

Enzyme Sequence structure in PDB —> https://www.ebi.ac.uk/pdbe/entry/pdb/1QQ0

Each letter corresponds to an amino acid, each amino acid has a different physical shape, function and unique physiochemical characteristics. Amino acids physically connect to form a chain and this long chain of amino acids is then folded in on itself to form a protein of a specific shape. By knowing the unique physiochemical characteristics of the amino acids and how they are arranged in the sequence algorithms know the shape and overall physiochemical characteristics of the protein. Protein Data Bank has all the known shapes and overall characteristics mapped out in great detail, including for this enzyme. *enzymes are just proteins that do something* The pic on the right is the 3D folded image of the series of letters.



Enzyme code _ EC:4.2.1.1
https://en.wikipedia.org/wiki/Enzyme_Commission_number

Every enzyme code consists of the letters "EC" followed by four numbers separated by periods. The PDB assigns EC (Enzyme Commission) numbers to proteins according to the type of chemical reaction catalyzed, specific donors and receptors of chemical groups participating in the reactions etc. This makes it simple to know what the enzyme does and serves as its classifier.

Example: Search for structures of DNA polymerase, an enzyme responsible for copying genetic material to make new copies of DNA, can be done as follows:

- browse the tree for “Transferases” >> “Transferring phosphorus-containing groups” >> “Nucleotidyltransferases” and “DNA-directed DNA polymerase” OR
- search by typing DNA polymerase in the search box on the page and select from the options “DNA-directed DNA polymerase” or typing in the EC number (type 2.7.7.7)

A mutation in one can lead to the enzyme functioning better, worse, or the same. For Example:

EC:4.2.1.1
Assay activity for the breakdown of borosilicate clay
We mutate the enzymes and run a test to see how well each one broke down the specific rock

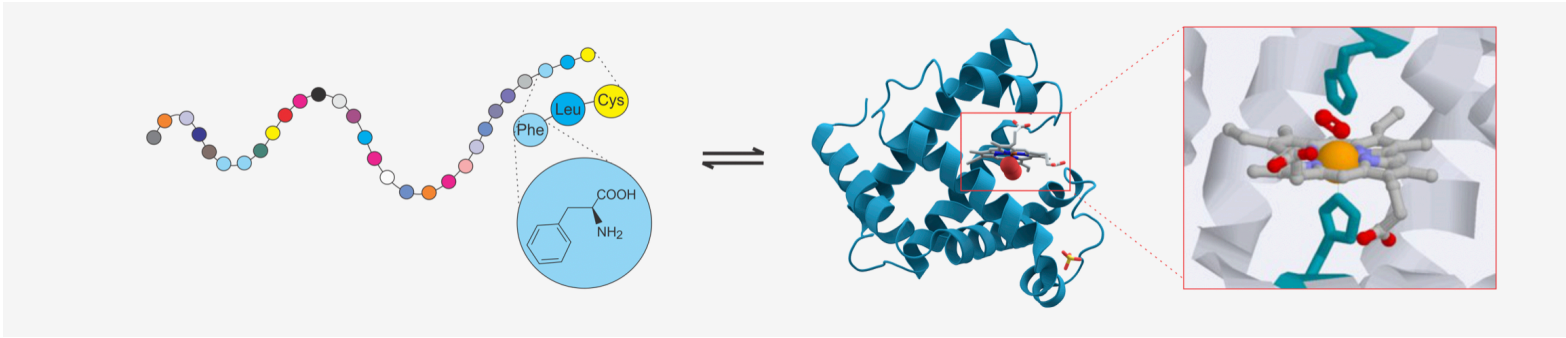
Protein sequence 0 (no Mutation) Activity Score (5-baseline)
MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNVQDGVVLHALETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSQVHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTNEAVVYVNVHLAEGYKETS

Protein sequence 1 (Mutation) Activity Score (4)
MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNVQDGVVLHALETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSLVHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTNEAVVYVNVHLAEGYKETS

Protein sequence 2 (Mutation) Activity Score (7)
MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGGRSNVQDGVVLHALETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSQVHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTNEAVVYVNVHLAEGYKETS

Protein sequence 3 (Mutation) Activity Score (10)
MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNVQDGVVLIYLETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSQVHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTNEAVVYVNVHLAEGYKETS

In this case, mutant 1 performed worse than the baseline, and mutants 2 and 3 did better at breaking down the specific borosilicate clay under a set of conditions. This is because a change in the amino acid sequence (letter) changed the physiochemical characteristics of the 3D protein overall. Maybe this made it harder for the enzyme to hold onto the borosilicate clay in mutant 1, or maybe it made it easier in some way in mutants 2&3.



The ML model comes in and learns what those specific changes are in the physiochemical characteristics of the 3D protein. (Did it get more positive, smaller, more acidic, etc) It does so be being trained by the 3D protein data available in PDB. When we input our data and classify the enzyme a particular type, it works its magic and can understand what changes were made and potentially what changes it would recommend, based on what worked for the real data and what it learned from the PDB. In this case it would be a generative model.

Protein sequence 4 (Recommended Mutation) Predicted Activity Score (15)
MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNVQDGVVLIYLETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSIRHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTEEAVVYVNVHLAEGYKETS

To test this, we would print the DNA, put it into a bacteria, and then test the enzyme it makes using the same assay as we had been using.

Protein sequence 4 (Recommended Mutation) Actual Activity Score (15.005)
MMFNKQIFTILILSLSLALAGSGCISEGAEDNVAQEITVDEFSNIRENPVTPWNPEPSAPVIDPTAYI
DPQASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNVQDGVVLIYLETINEEGEPIEDNIVEV
DGKEYAVYIGNNVSLAHQSIRHGPAAVGDDTFIGMQAFVFKSKVGNNCVLEPRSAAIGVTIPDGR
YIPAGMVVTSQAEADKLPEVTDDYAYSHTEEAVVYVNVHLAEGYKETS

In this case, the model was spot on.

We want to tackle different iterations of the enzyme for different rock samples. For example, we could use it to generate the best-performing mutant for borosilicate clay and when we are done we then begin experiments to test mutants in Aluminosilicate clay. We would therefore need a model that we can train, and then say, open another browser window and re-train it with the same enzyme, but reacting towards a different rock.

The ideal input would be a CSV file with the sequence and the experimental activity score along with the 4-digit classifier code.