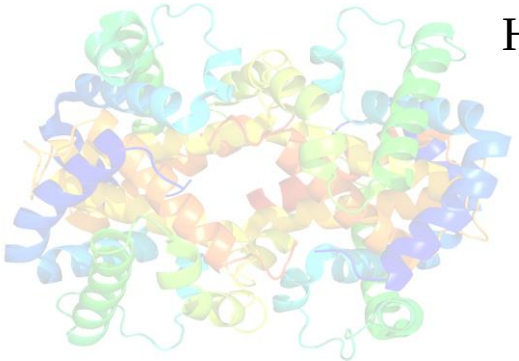
PDB AUTOFILL

Technology Review

Henry Lee, You-Hsin Chen, Jenny Bennett



RCSB PDB
PROTEIN DATA BANK

UNIVERSITY of WASHINGTON

Background

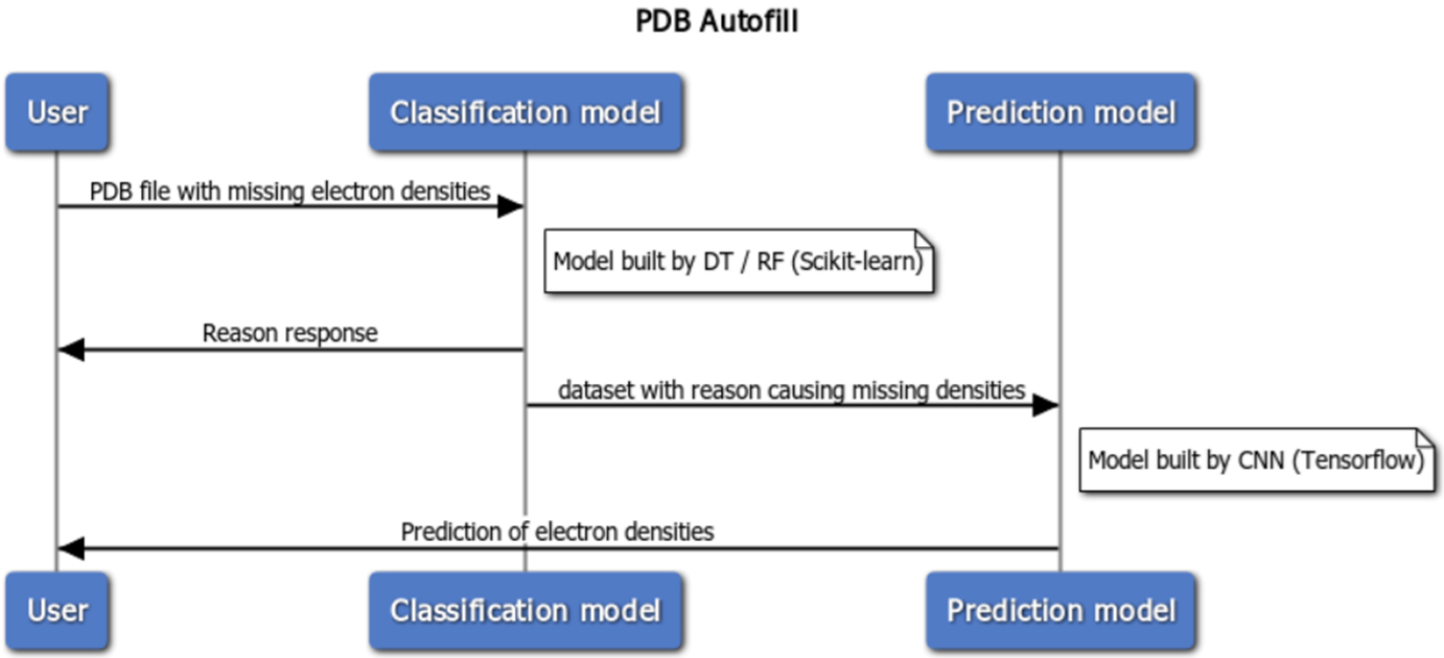
- From the Protein Data Bank website: Many PDB entries are missing portions of the molecule that were not observed in the experiment
- Our goal: Classify the reason for missing electron densities of protein crystals in PDB by using a decision tree model
- Stretch: Predict the missing density coordinates in proteins with a neural network model
- Entity 1AZ5 skips residues 48-51

Sequence Number

```
ATOM    437  CG2 VAL A  47
ATOM    438  H   VAL A  47
ATOM    439  N   GLY A  52
ATOM    440  CA  GLY A  52
```

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     GLY A    48
REMARK 465     GLY A    49
REMARK 465     ILE A    50
REMARK 465     GLY A    51
REMARK 470
REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS (M=MODEL NUMBER;
REMARK 470 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 470 I=INSERTION CODE):
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     PHE A  53    CG   CD1  CD2  CE1  CE2  CZ
REMARK 470     LYS A  60    CG   CD   CE   NZ
```

Diagram



UNIVERSITY of WASHINGTON

Seaborn

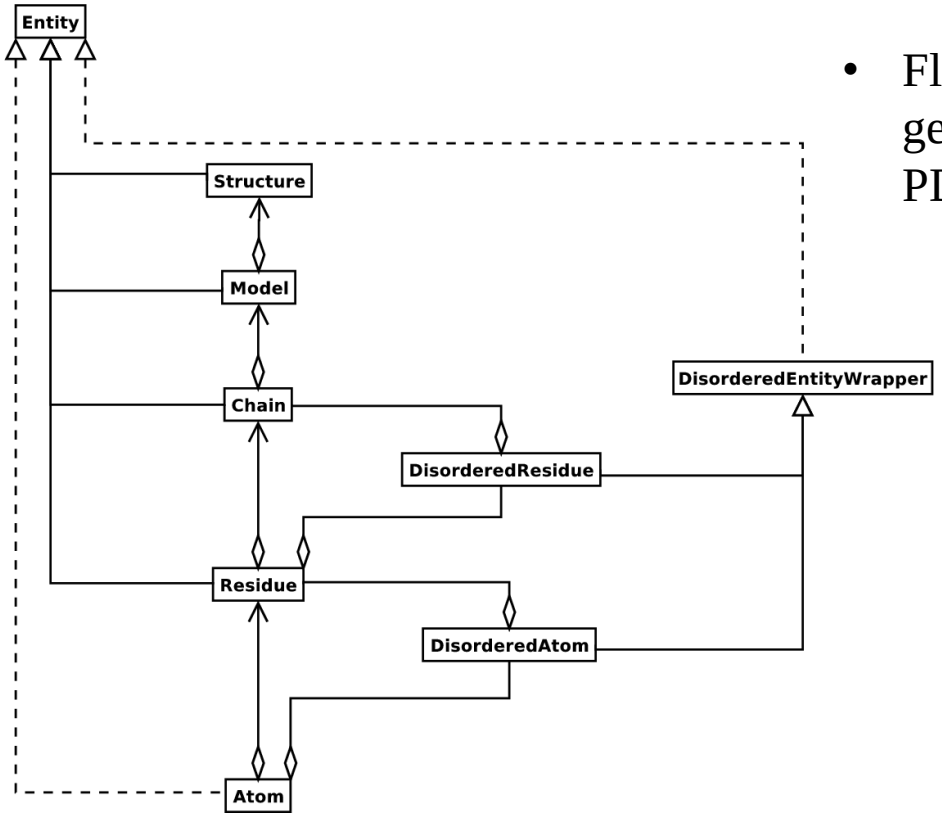
- Background
  - Plotting for visualization of our classification
- How the package works
  - It is built on top of matplotlib, but provides attractive and informative statistical graphs
- Appeal of using the package
  - Works well with pandas dataframes
  - Comes with a large number of customized themes
  - Simple to use
- Drawbacks of using the package
  - There are higher quality visualization packages such as Plotly, which are more complicated and unfamiliar to use

Biopython

- Background
  - Extract data from PDB file.
- How the package works
  - It’s a set of useful tools for biological computation
  - In our project, we will use Bio.PDB for getting structures of proteins from Protein Data Bank (PDB).
- Appeal of using the package
  - It’s designed for biological using, so it has many commands which can help us parse data from PDB without long code.
  - Different parts can interoperate (no reading and writing of file format)
- Drawbacks of using the package
  - Sometimes users have to learn the specific code to do the way you want (e.g. extract residue name or residue id)

UNIVERSITY of WASHINGTON

Biopython



- Flowchart for how Biopython get structure of protein in PDB file

UNIVERSITY of WASHINGTON

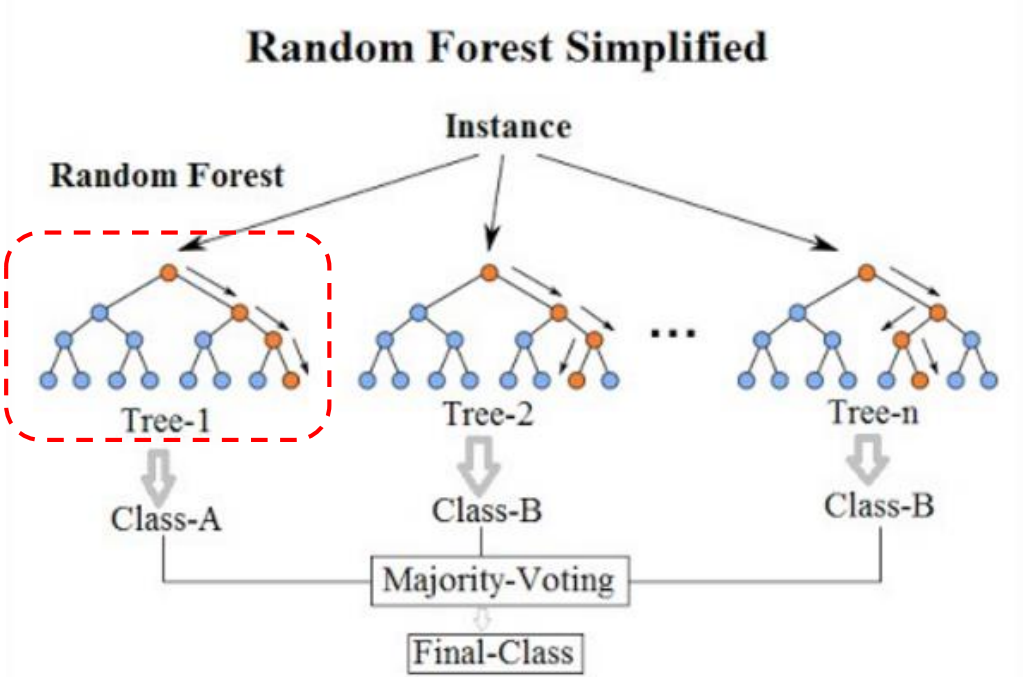
Scikit-learn

- Background
  - Build up model for classifying the reasons which cause missing densities.
- How the package works
  - It's a great package which has comprehensive functions in machine learning like classification, regression, clustering …etc.
  - In our project, we will use Decision Tree and Random Forest to determine what resulting in missing densities.
  - Also, we would utilize functions in it to list importance of each parameter.
- Appeal of using the package
  - Scikit-Learn has good documentation, and works well with Numpy.
  - We can execute complicated algorithm without writing the whole math derivation and formula by ourselves.
- Drawbacks of using the package
  - Sometimes the package lacks of flexibility.

UNIVERSITY of WASHINGTON

Scikit-learn

Decision Tree



UNIVERSITY of WASHINGTON

Tensorflow

- Background
  - Prediction of electron densities (missing atom positions)
- How the package works
  - Takes input as tensors and operates them through dataflow graphs.
- Appeal of using the package
  - There are a lot of tutorials for tensorflow.
  - Has built-in visualization tool (TensorBoard)
  - Provides data pipeline tools which handle complex input data.
- Drawbacks of using the package
  - It’s hard to figure out error messages, which makes the debugging process difficult.

UNIVERSITY of WASHINGTON