# README

## Overview

This Python script processes Protein Data Bank (PDB) files to generate **3D molecular fingerprints**. These fingerprints are rich in information and include:

* Functional groups (e.g., hydrogen bond donors, carboxylic acids, etc.).
* Molecular scaffolds.
* Physical properties like molecular mass, polar surface area, and rotatable bonds.

The fingerprints are compiled into a pandas DataFrame (fpArray) which can be used as input for machine learning algorithms.

## Features

1. **Batch Processing of PDB Files**:
   * Iteratively processes all PDB files in a specified directory (D:\Professional\TheProject\Codes\Structures\pdb\).
2. **Physical Property Calculation**:
   * Molecular mass.
   * Octanol-water partition coefficient (MLogP).
   * Topological polar surface area (TPSA).
   * Number of rotatable bonds.
   * Hydrogen bond acceptors and donors.
3. **Functional Group Identification**:
   * Detects and quantifies functional groups such as:
     + Carboxylic acids.
     + Carboxylates.
4. **SMILES Conversion**:
   * Converts PDB files to SMILES strings for additional molecular property computation.
5. **Detailed 3D Molecular Data**:
   * Calculates distances and angles between atoms in 3D space.
   * Identifies overlapping and interacting atomic groups.
6. **Output**:
   * The final DataFrame, fpArray, is written to a CSV file (fpArray.csv) for downstream analysis.

## Requirements

* Python 3.x
* Required packages:
  + numpy
  + pandas
  + tqdm
  + rdkit
  + biopandas
  + sympy
  + biopython

Install missing dependencies using:

bash

pip install numpy pandas tqdm rdkit biopandas sympy biopython

## Usage

### Input

1. **PDB Files**:
   * Place your PDB files in the directory D:\Professional\TheProject\Codes\Structures\pdb\.
   * Filenames should follow the format <molecule\_number>.pdb.
2. **Customizing File Path**:
   * If your files are stored in a different location, update the fpath variable in the script.

### Execution

1. Run the script.
2. The function molFp(mNum) will:
   * Parse the specified PDB file.
   * Calculate the molecular properties and functional groups.
   * Append the resulting fingerprint to fpArray.
3. After processing all PDB files, the script writes the fpArray DataFrame to a CSV file (fpArray.csv).

## Example

To process a molecule with the number 123:

python

fpnames = molFp(123)

To process all molecules in the directory:

bash

python your\_script\_name.py

## Notes

* **Functional Group Detection**:
  + Hydrogen bond donors are identified based on proximity criteria (e.g., H atoms within 0.5–1.2 Å of electronegative atoms like N, O, or F).
  + Carboxylic acids and carboxylates are identified using specific interatomic distance thresholds.
* **Logging**:
  + The script outputs the count of processed files for monitoring progress.
* **Error Handling**:
  + The script attempts to handle missing or improperly formatted files gracefully.