Dataset Analysis of Protein Structures

Dataset Overview

In this notebook, we load a curated list of PDB identifiers, download their corresponding structures, and compute the following metrics for each protein:

- **Residue count**: total number of amino acid residues.
- **Chain count**: number of polypeptide chains.
- **Center of mass**: 3D coordinates of the center of mass (X, Y, Z).
- **Average backbone angles**: mean φ (phi) and ψ (psi) angles in degrees.

In [13]:

```
%pip install -r ../requirements.txt
```

```
Requirement already satisfied: biopython>=1.79 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-
packages (from -r ../requirements.txt (line 1)) (1.85)
Requirement already satisfied: matplotlib>=3.5.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\sit
e-packages (from -r ../requirements.txt (line 2)) (3.10.3)
Requirement already satisfied: numpy>=1.23.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pa
ckages (from -r ../requirements.txt (line 3)) (2.2.6)
Requirement already satisfied: flask>=3.1.1 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pac
kages (from -r ../requirements.txt (line 4)) (3.1.1)
Requirement already satisfied: gunicorn>=20.1.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site
-packages (from -r ../requirements.txt (line 5)) (23.0.0)
Requirement already satisfied: pytest>=7.0.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pa
ckages (from -r ../requirements.txt (line 6)) (8.3.5)
Requirement already satisfied: requests>=2.28.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site
-packages (from -r ../requirements.txt (line 7)) (2.32.4)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site
-packages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (1.3.2)
Requirement already satisfied: cycler>=0.10 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pac
kages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\sit
e-packages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (4.58.1)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\sit
e-packages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (1.4.8)
Requirement already satisfied: packaging>=20.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-
packages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (25.0)
Requirement already satisfied: pillow>=8 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-packag
es (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (11.2.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site
-packages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (3.2.3)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\
site-packages (from matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (2.9.0.post0)
Requirement already satisfied: blinker>=1.9.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-p
ackages (from flask>=3.1.1->-r ../requirements.txt (line 4)) (1.9.0)
Requirement already satisfied: click>=8.1.3 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pac
kages (from flask>=3.1.1->-r ../requirements.txt (line 4)) (8.2.1)
Requirement already satisfied: itsdangerous>=2.2.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\s
ite-packages (from flask>=3.1.1->-r ../requirements.txt (line 4)) (2.2.0)
Requirement already satisfied: jinja2>=3.1.2 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pa
ckages (from flask>=3.1.1->-r ../requirements.txt (line 4)) (3.1.6)
Requirement already satisfied: markupsafe>=2.1.1 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\sit
e-packages (from flask>=3.1.1->-r ../requirements.txt (line 4)) (3.0.2)
Requirement already satisfied: werkzeug>=3.1.0 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-
packages (from flask>=3.1.1->-r ../requirements.txt (line 4)) (3.1.3)
Requirement already satisfied: colorama in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-package
s (from pytest>=7.0.0->-r ../requirements.txt (line 6)) (0.4.6)
Requirement already satisfied: iniconfig in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-packag
es (from pytest>=7.0.0->-r ../requirements.txt (line 6)) (2.1.0)
Requirement already satisfied: pluggy<2,>=1.5 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-p
ackages (from pytest>=7.0.0->-r ../requirements.txt (line 6)) (1.6.0)
Requirement already satisfied: charset_normalizer<4,>=2 in c:\users\user\pycharmprojects\protein_explorer\.venv\
lib\site-packages (from requests>=2.28.0->-r ../requirements.txt (line 7)) (3.4.2)
Requirement already satisfied: idna<4,>=2.5 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-pac
kages (from requests>=2.28.0->-r ../requirements.txt (line 7)) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\si
te-packages (from requests>=2.28.0->-r ../requirements.txt (line 7)) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\si
te-packages (from requests>=2.28.0->-r ../requirements.txt (line 7)) (2025.6.15)
Requirement already satisfied: six>=1.5 in c:\users\user\pycharmprojects\protein_explorer\.venv\lib\site-package
s (from python-dateutil>=2.7->matplotlib>=3.5.0->-r ../requirements.txt (line 2)) (1.17.0)
Note: you may need to restart the kernel to use updated packages.
```

In [14]:

```
import os
import pandas as pd
import numpy as np
```

```
from Bio.PDB import PDBList, PDBParser
import plotly.express as px

pdb_list_file = '../data/pdb_list.txt'
```

In [15]:

```
with open(pdb_list_file) as f:
    pdb_ids = [line.strip() for line in f if line.strip()]

pdb_ids[:10]
```

Out[15]:

```
['1AKE', '1PKE', '2PTC', '2PTK', '5XNL', '6LU7', '7AHD']
```

In [16]:

```
from Bio.PDB.vectors import calc_dihedral
from Bio.PDB import PPBuilder

pdbl = PDBList()
parser = PDBParser(QUIET=True)

os.makedirs('structures', exist_ok=True)


def download_structure(pdb_id, out_dir='structures'):
    path = pdbl.retrieve_pdb_file(pdb_id, pdir=out_dir, file_format='pdb')
    return path

def parse_structure(path):
    return parser.get_structure(path, path)

def compute_center_of_mass(struct):
    coords = []
    masses = []
    for atom in struct.get_atoms():
        if hasattr(atom, 'mass'):
            coords.append(atom.get_coord() * atom.mass)
            masses.append(atom.mass)
    if not masses:
        return (float('nan'), float('nan'), float('nan'))
    com = np.sum(coords, axis=0) / np.sum(masses)
    return tuple(com)

def compute_phi_psi(struct):
    angles = []
    for model in struct:
        for chain in model:
            for poly in PPBuilder().build_peptides(chain):
                for phi, psi in poly.get_phi_psi_list():
                    if phi and psi:
                        angles.append((phi, psi))
    return angles
```

In [17]:

```
results = []
for pdb_id in pdb_ids:
    path = download_structure(pdb_id)
    struct = parse_structure(path)
    residues = [res for res in struct.get_residues() if res.id[0] == ' ']
    residue_count = len(residues)
    chain_count = len({chain.id for chain in struct.get_chains()})
    center_of_mass = compute_center_of_mass(struct)
    phi_psi = compute_phi_psi(struct)
    if phi_psi:
        avg_phi = np.degrees(np.mean([a for a,_ in phi_psi]))
        avg_psi = np.degrees(np.mean([b for _,b in phi_psi]))
    else:
        avg_phi = float('nan')
        avg_psi = float('nan')
    results.append({'pdb_id': pdb_id, 'residue_count': residue_count, 'chain_count': chain_count,
                    'center_x': center_of_mass[0], 'center_y': center_of_mass[1], 'center_z': center_of_mass[2]
                    'avg_phi': avg_phi, 'avg_psi': avg_psi})

df = pd.DataFrame(results)
print(df.head())
print(df.describe())
```

```
Structure exists: 'structures\pdb1ake.ent'
Structure exists: 'structures\pdb1pke.ent'
Structure exists: 'structures\pdb2ptc.ent'
Structure exists: 'structures\pdb2ptk.ent'
Structure exists: 'structures\pdb5xnl.ent'
Structure exists: 'structures\pdb6lu7.ent'
Structure exists: 'structures\pdb7ahd.ent'
  pdb_id  residue_count  chain_count   center_x   center_y   center_z  \
0   1AKE            428            2  20.184478  25.612156  20.331380   
1   1PKE            706            3  76.646998  45.124234  19.631075   
2   2PTC            281            2  10.865600  70.390938  17.345945   
3   2PTK            425            1  24.063875  -0.259675  35.190393   
4   5XNL           9364           56   0.003145   0.003233 -35.104994   

     avg_phi    avg_psi  
0 -68.610796  25.445530  
1 -79.067611  40.456458  
2 -78.083670  67.282326  
3 -75.165537  43.502933  
4 -71.330482   9.501080  
       residue_count  chain_count    center_x    center_y    center_z  \
count       7.000000     7.000000    7.000000    7.000000    7.000000   
mean     1839.142857    10.000000   34.010630   40.589842   34.868396   
std      3339.071200    20.305993   53.350017   47.160672   49.696432   
min       281.000000     1.000000  -26.041303   -0.259675  -35.104994   
25%       367.000000     2.000000    5.434373    6.301372   18.488510   
50%       428.000000     2.000000   20.184478   25.612156   20.331380   
75%      1033.500000     3.500000   50.355436   57.757586   47.189197   
max      9364.000000    56.000000  132.351617  130.658496  127.496970   

         avg_phi    avg_psi  
count   7.000000   7.000000  
mean  -74.702465  35.295484  
std     4.781010  23.308029  
min   -80.676619   4.624284  
25%   -78.575641  17.473305  
50%   -75.165537  40.456458  
75%   -70.656512  49.879355  
max   -68.610796  67.282326  
```

Center of Mass Scatter (X vs Y)

This scatter plot displays the X and Y coordinates of each protein’s center of mass.

- Each point represents one PDB structure.
- Color encodes the number of chains in that structure (warmer colors = more chains).

By examining this plot, we can see how protein mass is distributed spatially (in the XY plane) and whether multi-chain complexes tend to occupy different spatial regions than single-chain proteins.

Center of Mass Scatter (X vs Y)

This scatter plot displays the X and Y coordinates of each protein’s center of mass.

- Each point represents one PDB structure.
- Color encodes the number of chains in that structure (warmer colors = more chains).

By examining this plot, we can see how protein mass is distributed spatially (in the XY plane) and whether multi-chain complexes tend to occupy different spatial regions than single-chain proteins.

Average Phi and Psi Angle Distributions

Here we compare the backbone dihedral angles across all proteins:

- **avg_phi** box shows the distribution of mean φ angles.
- **avg_psi** box shows the distribution of mean ψ angles.

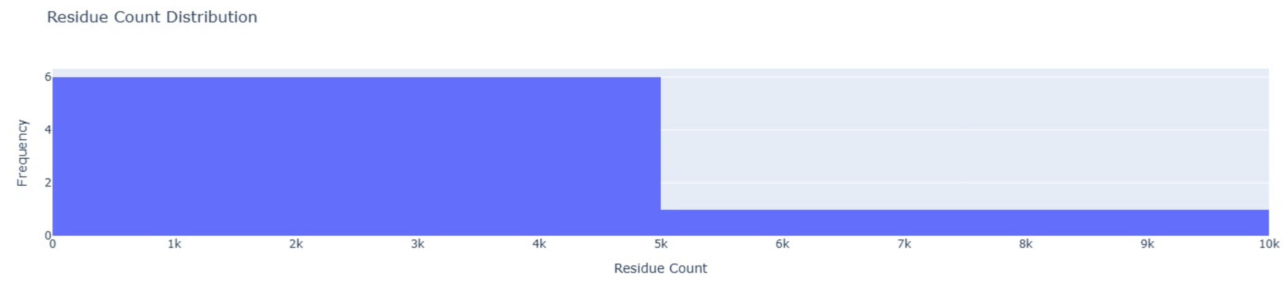
Box plots summarize the median, interquartile range, and outliers for each angle.
This analysis provides insight into the typical backbone conformations in our dataset.

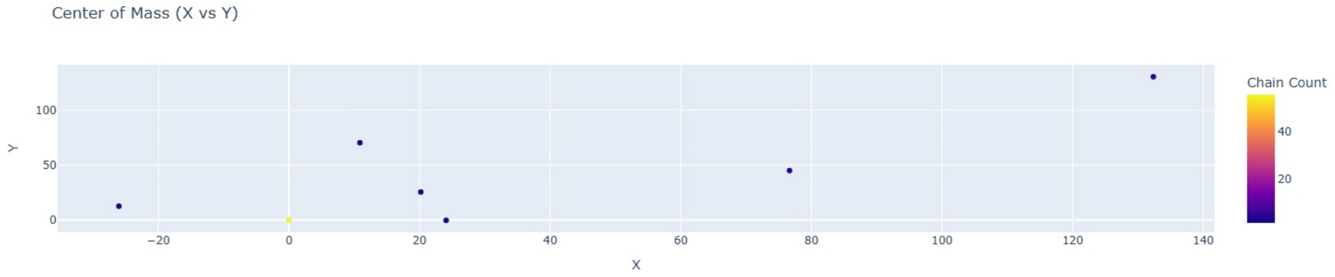
In [5]:

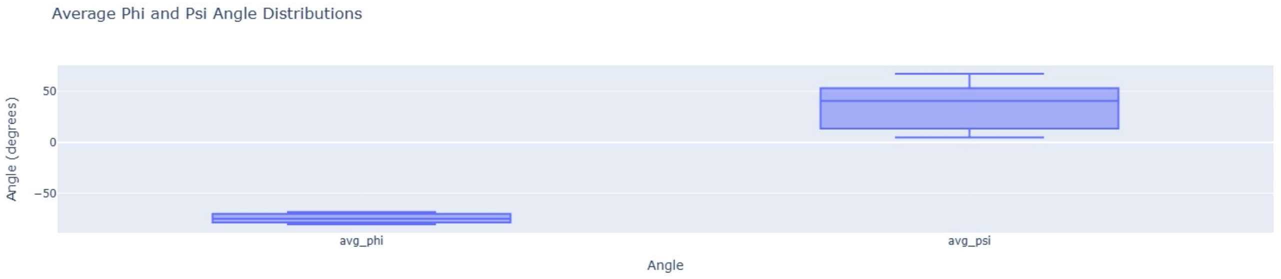
```
fig1 = px.histogram(df, x='residue_count', title='Residue Count Distribution')
fig1.update_xaxes(title='Residue Count')
fig1.update_yaxes(title='Frequency')
fig1.show()

fig2 = px.scatter(df, x='center_x', y='center_y', color='chain_count',
                  title='Center of Mass (X vs Y)', labels={'center_x':'X', 'center_y':'Y', 'chain_count':'Chain (
fig2.show()
```

```
fig3 = px.box(df, y=['avg_phi', 'avg_psi'], title='Average Phi and Psi Angle Distributions',
             labels={'value':'Angle (degrees)', 'variable':'Angle'})
fig3.show()
```







Exporting Metrics to CSV

Finally, we save the compiled metrics into `results/dataset_metrics.csv`, which can be used for further downstream analysis or shared with collaborators.

In [19]:

```
os.makedirs('results', exist_ok=True)
df.to_csv('results/dataset_metrics.csv', index=False)
```

In [ ]: