Medvolt Tech Private Limited

Assignment: Feature Engineering for Structure-Based Generative Learning

Background

In structure-based generative drug discovery, the quality of representations derived from protein structures directly impacts downstream learning tasks. Before model selection or training, robust feature extraction and encoding pipelines are critical to ensure that structural, chemical, and geometric information is captured in a machine-readable and generative-friendly format.

This assignment focuses exclusively on **feature extraction, representation design, and encoding-decoding logic**, not on model architecture or training.

Problem Statement

You are provided with a **protein structure in PDB format**.

Your task is to design and implement a **custom feature extraction and encoding-decoding pipeline** suitable for use in a **generative deep learning workflow**.

Specifically, you must:

1. **Parse the PDB file** and extract meaningful structural and physicochemical information.

2. **Design a feature representation** that captures relevant biological, chemical, and spatial properties.

3. **Implement a Feature Encoder** that transforms the extracted features into a machine-learning-friendly format.

4. **Implement a Feature Decoder** that can reconstruct the original representation (or a well-defined approximation) from the encoded form.

You are **not required** to build, train, or evaluate any deep learning model.

Scope and Expectations

1. Feature Extraction

You may choose the level of representation, but it must be clearly justified. Examples include but are not limited to:

- Atom-level features
- Residue-level features
- Graph-based representations
- Geometric or spatial descriptors

Feature choices may include:

- Atom or residue types
- Coordinates and distances
- Bond or neighborhood information
- Secondary structure or local environment descriptors
- Physicochemical properties

Explain why your chosen features are suitable for a **generative modeling context**.

2. Feature Encoding

Design an encoder that:

- Converts raw extracted features into a structured numerical representation
- Is deterministic and reproducible
- Can serve as input to a generative deep learning model

Examples:

- Graph encodings
- Tensor representations
- Latent or compressed feature formats
- Structured dictionaries with defined schemas

3. Feature Decoding

Design a decoder that:

- Reconstructs the original feature representation or a well-defined reduced form
- Preserves essential structural or chemical information
- Demonstrates reversibility or interpretability of the encoding

Exact atomic reconstruction is not mandatory, but the decoding objective must be clearly defined.

Deliverables

1. **Code Implementation**
    - Feature extractor
    - Encoder
    - Decoder
    - Clear modular structure
2. **Documentation**
    - Explanation of feature choices
    - Description of encoding and decoding logic
    - Assumptions and design trade-offs
3. **Optional Bonus**
    - Visualization of extracted features
    - Validation checks for encoding-decoding consistency
    - Discussion on scalability to large protein datasets

Evaluation Criteria

You will be evaluated on:

- Quality and relevance of feature design
- Clarity of reasoning and documentation
- Code structure, readability, and robustness
- Understanding of structural biology and ML representation challenges
- Practical suitability for generative modeling workflows

Notes

- You may use standard Python libraries such as NumPy, PyTorch, RDKit, or BioPython.
- Focus on **engineering rigor and representation design**.
- Creativity and clear justification are strongly encouraged.