Clustering and Feature Analysis of HIV Active Molecules

Background and Research Objectives

This research consists of two main parts:

- Regression analysis, which aims to predict the molecular bioactivity levels using supervised learning methods, thereby establishing a quantitative relationship between physicochemical properties and anti-HIV activity (conducted by other members of the project team).
- Clustering analysis (focus of this report), which applies unsupervised learning approaches based on molecular descriptors to identify intrinsic data patterns. This study employs classical unsupervised clustering algorithms such as PCA + DBSCAN, as well as a deep feature learning model based on an Autoencoder, to systematically analyze the clustering patterns of HIV molecular bioactivity data.

By combining molecular descriptors and structure-based fingerprints, we aim to elucidate the underlying relationship between chemical structure and bioactivity, providing valuable insights for drug design, virtual screening, and molecular optimization.

Experiment 1-PCA+DBSCAN

1.Data Preprocessing

The original dataset contains approximately 42,000 SMILES molecular sequences, each labeled with its corresponding HIV bioactivity status (Active / Inactive).
Prior to clustering analysis, a comprehensive data preprocessing procedure was carried out to ensure structural validity and consistency:

- Removal of null or unlabeled entries, ensuring that each molecule retained both valid structural information and its bioactivity annotation;
- Filtering of invalid SMILES strings that could not be parsed by RDKit.
  Such invalid entries typically resulted from syntax errors or chemically unreasonable valence states

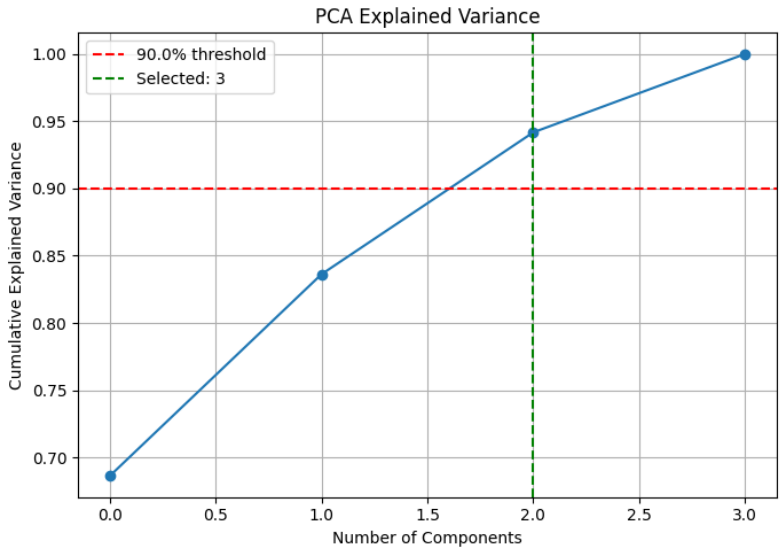
After preprocessing, approximately 41,287 valid molecules remained in the dataset, providing a reliable foundation for feature extraction and clustering analysis.

2. Feature Extraction

After data preprocessing, molecular samples were transformed into numerical feature representations and normalized for clustering analysis.
Two distinct feature extraction approaches were adopted, corresponding to different clustering strategies:

- Structure-based Clustering
  Each SMILES string was converted into a molecular fingerprint, and molecular similarity was quantified using the Tanimoto distance.
  This representation captures structural similarity at the scaffold level, enabling clustering based on chemical topology.
- Property-based Clustering
  A comprehensive set of molecular descriptors was computed using RDKit, including molecular weight (MolWt), logarithm of the partition coefficient (LogP), topological polar surface area (TPSA), hydrogen-bond donors/acceptors (HBD/HBA), and the number of rotatable bonds[1].
  This approach focuses on clustering according to physicochemical "activity patterns," aiming to identify properties correlated with HIV bioactivity.

Subsequently, Principal Component Analysis (PCA) was applied for dimensionality reduction.
Among approximately 200 molecular descriptors, the first three principal components explained over 95% of the total variance, while PC1 and PC2 alone captured more than 90%. Thus, all clustering analyses were performed in the 2D PCA space, effectively reducing computational cost and mitigating feature noise.



3.Clutering based on Classic Machine Learning Model

3.1 Structure-based Clustering (Morgan Fingerprint + Tanimoto + DBSCAN)

In the traditional machine learning experiment, a structure-based clustering approach was first attempted.
Each molecule's SMILES representation was converted into a Morgan fingerprint, and pairwise Tanimoto similarities were computed and transformed into a distance matrix for DBSCAN clustering[2].

However, this method faced severe computational limitations. With approximately 42,000 molecules and 1024-bit fingerprints, constructing the full Tanimoto distance matrix required over 1.6 billion pairwise calculations ($O(n^2)$), consuming tens of gigabytes of memory and resulting in extremely low efficiency. A smaller subset of about 2,000 molecules was tested, yet the results remained unsatisfactory: around 97% of samples were classified as noise points (Cluster = –1), with only a few small dense clusters detected.

In summary, the Fingerprint + Tanimoto + DBSCAN combination is unsuitable for large-scale molecular clustering, mainly due to feature sparsity and computational complexity.

3.2 DBSCAN Clustering Based on Molecular Descriptors and Parameter Optimization

3.2.1 Parameter Selection and Clustering Results

After testing multiple parameter combinations, the highest silhouette score (0.60) was achieved at eps = 0.60 and min_samples = 5, which were selected as the optimal settings.

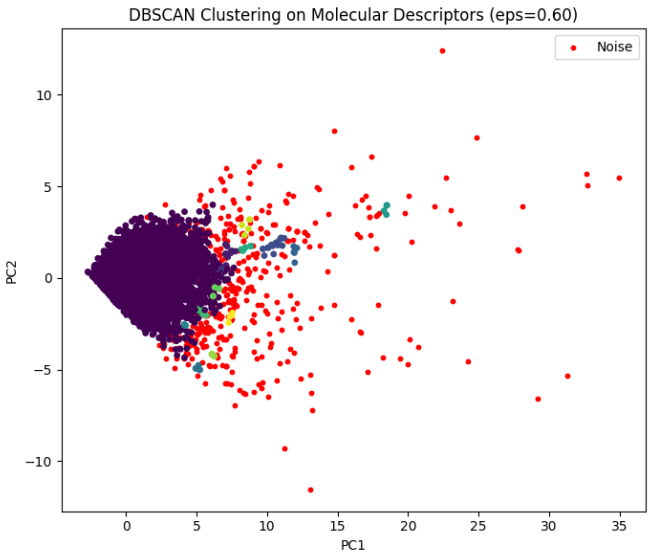
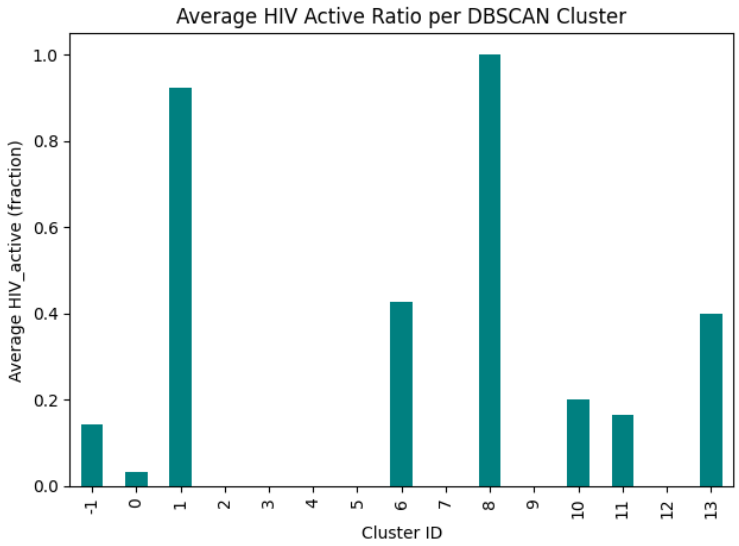


Figure DBSCAN clustering based on molecular descriptors (eps = 0.60). Molecular physicochemical descriptors were extracted using RDKit, and DBSCAN clustering was

performed in the PCA-reduced feature space. Points represent molecular samples projected on the first two principal components (PC1 and PC2), with red dots indicating noise points (label = –1).



Average HIV Activity per Cluster (Horizontal Layout)

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0.144	0.034	0.923	0.000	0.000	0.000	0.000	0.429	0.000	1.000	0.000	0.200	0.167	0.000	0.400

Table shows the average HIV activity ratio for each cluster identified by DBSCAN. Cluster 1 (0.923) and Cluster 8 (1.000) exhibit the highest activity enrichment.

3.2.2 Results Analysis and Visualization

- Representative Scaffolds

Both high-activity clusters (Cluster 1 and Cluster 8) share similar structural motifs. Their representative scaffolds feature multiple amide linkages (–CO–NH–) and polycyclic aromatic frameworks, indicating large, planar, and conjugated molecular structures. Such characteristics contribute to enhanced hydrogen bonding interactions with biological targets, explaining their higher observed bioactivity.

- Functional Group Distribution

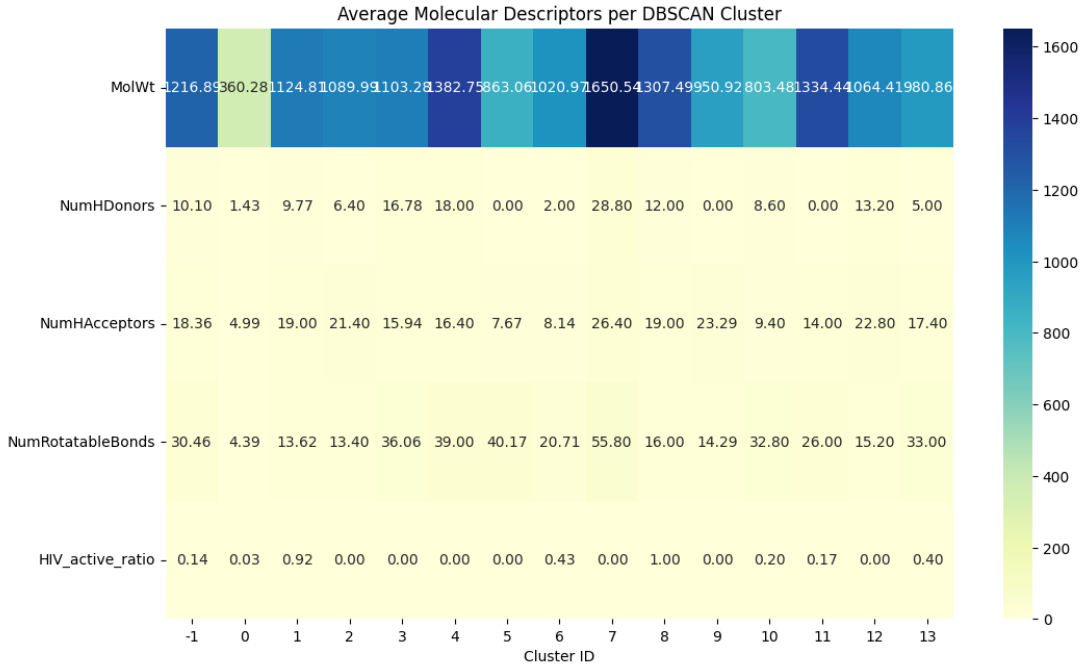
Functional Group Distribution

Functional group	Cluster 1 (%)	Cluster 8 (%)	Difference
Amide	**69.2**	**100**	**+30.8**
Amine	**23.1**	**20.0**	**-3.1**
Phenyl	**100**	**100**	**0**
Alcohol	**7.7**	**0**	**-7.7**
Phenol	**38.5**	**20.0**	**-18.5**

Both clusters exhibit abundant amide linkages, with Cluster 8 reaching 100% amide presence, indicating that all molecules contain amide bonds.

Cluster 1 also shows a high amide ratio but includes minor amounts of alcohol and phenol groups. This suggests that Cluster 8 compounds are more polar, potentially providing additional hydrogen-bonding sites and correlating with higher HIV inhibitory activity.

- Average Descriptor Comparison



Cluster 8 compounds exhibit larger molecular weights, stronger polarity, and higher numbers of hydrogen-bond donors (HBD) and rotatable bonds (RB).

This indicates a greater conformational flexibility and the potential for multipoint binding interactions, although it may also lead to reduced solubility.

Cluster 1 compounds also show high molecular weight and polarity but with slightly lower flexibility, suggesting more rigid and stable conformations that may favor target binding.

Overall, the high-activity clusters (Cluster 1 and Cluster 8) share several structural characteristics: Amide-rich scaffolds; Polycyclic aromatic frameworks; High polarity.

These common features suggest that highly active compounds tend to be planar, conjugated, and hydrogen-bond-rich, providing structural evidence for their enhanced bioactivity and potential binding affinity.

4.Problems and Prospect

Using DBSCAN clustering, two highly active molecular clusters (Cluster 1 and Cluster 8) were identified, with activity rates of 92.3% and 100%, far exceeding the baseline level. These clusters share peptide-like features—multiple amide linkages, polycyclic aromatic frameworks, and high polarity—closely resembling known HIV protease inhibitors. The findings provide structural guidance for HIV drug design, with Cluster 8's amide-rich, highly polar compounds meriting further experimental validation.

Future work will focus on:
(1) Handling data imbalance by applying sampling or density-based balancing methods to improve detection of active molecules;
(2) Optimizing fingerprint clustering through MiniBatch computation, approximate nearest neighbor search, or dimensionality reduction to enhance efficiency and scalability.

Experiment 2-AutoEncoder+HDBSCAN

1.Data Preprocessing

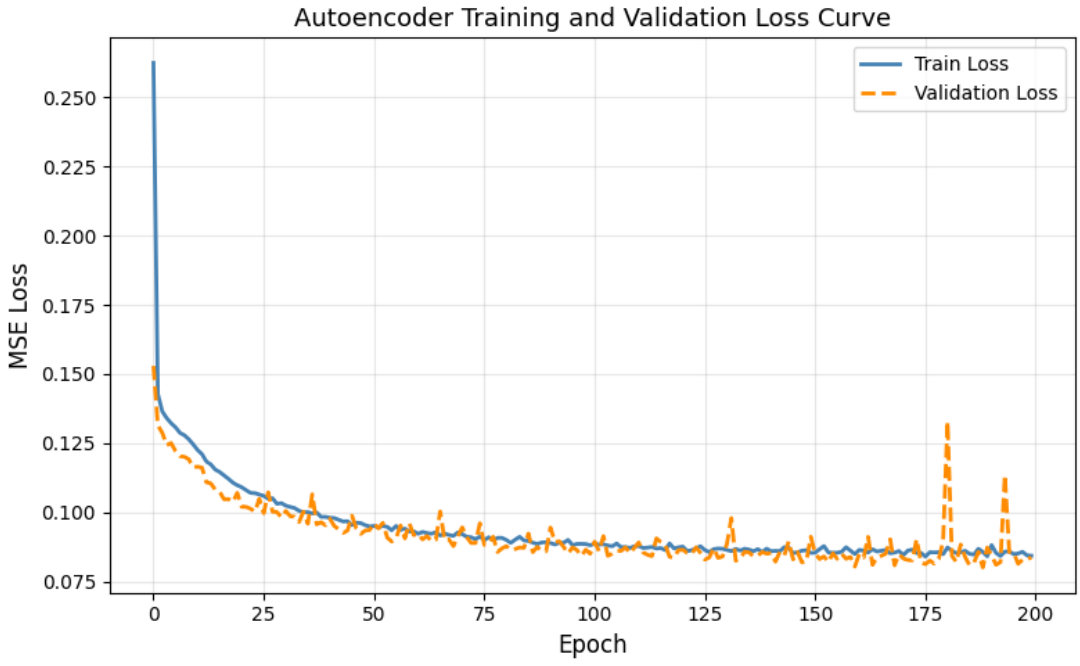
On the basis of Experiment 1, which focused on removing missing values and outliers, Experiment 2 adds data reduction and balancing functions. During testing, it was found that using the features extracted by the neural network for clustering could cause a memory overflow. Since this clustering mainly aims at identifying HIV-positive samples, the experiment extracts all positive samples and randomly matches them with three times as many negative samples (the ratio can be adjusted)[6]. The final positive input is 1443 and negative input is 4329.

2.Model Configuration and Training

This study employs the Mean Squared Error (MSE) as the loss function for the autoencoder. The MSE is chosen because the molecular descriptors are continuous numerical features, and it can directly measure the numerical difference between the reconstructed output and the original input. Using MSE ensures that the latent representations retain as much of the original molecular information as possible after compression[7].

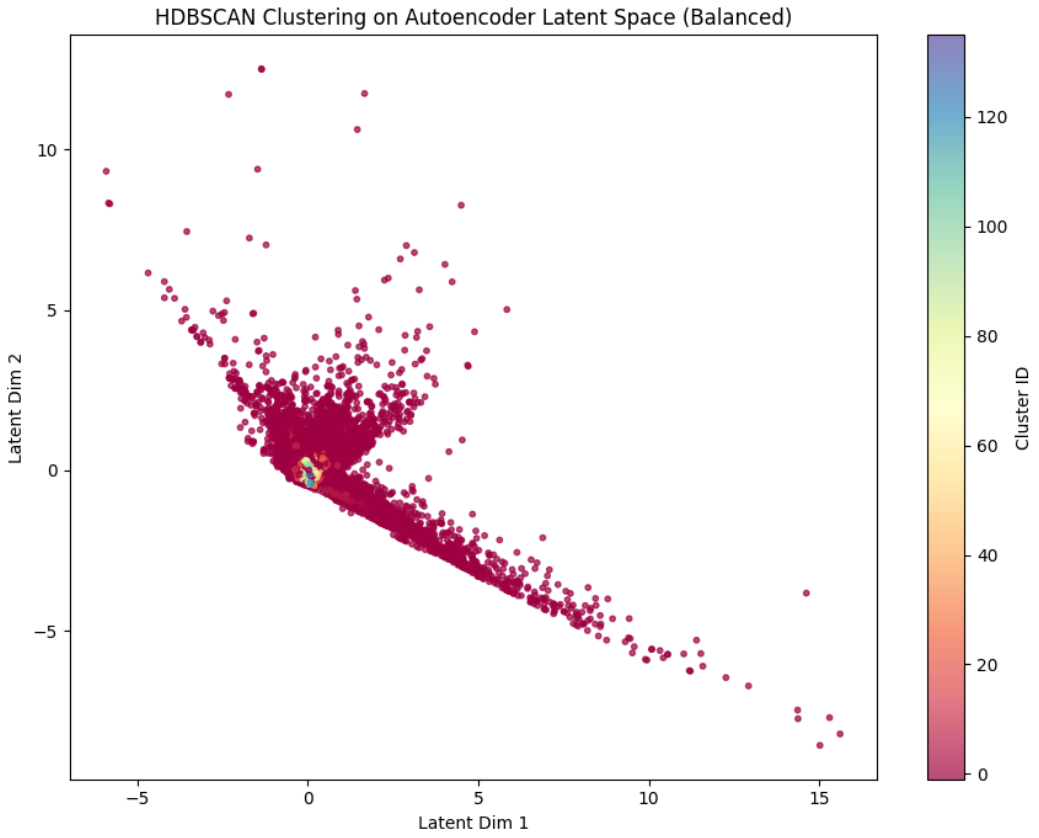
Parameter	Description	Value
Input size	Number of molecular descriptors	*dataset-dependent*
Hidden size	Neurons in hidden layer	32
Latent size	Dimensionality of latent space	2
Epochs	Training iterations	200
Learning rate	Optimization step size	1×10^{-3}
Loss function	Reconstruction loss	Mean Squared Error (MSE)
Optimizer	Gradient update algorithm	Adam

The following process shows the loss reduction of the model. It can be seen that when epoch=200, the model can basically reach the convergence state, Train Loss: 0.08426 | Eval Loss: 0.08374.



3. Autoencoder feature extraction

The model effectively compresses high-dimensional molecular descriptors into a low-dimensional latent space, enabling visualization and structural pattern extraction.
The latent distribution shows local clustering, indicating that the model captures molecular similarity. Combined with HDBSCAN, it automatically identifies potential molecular groups and supports subsequent structure–activity relationship analysis.



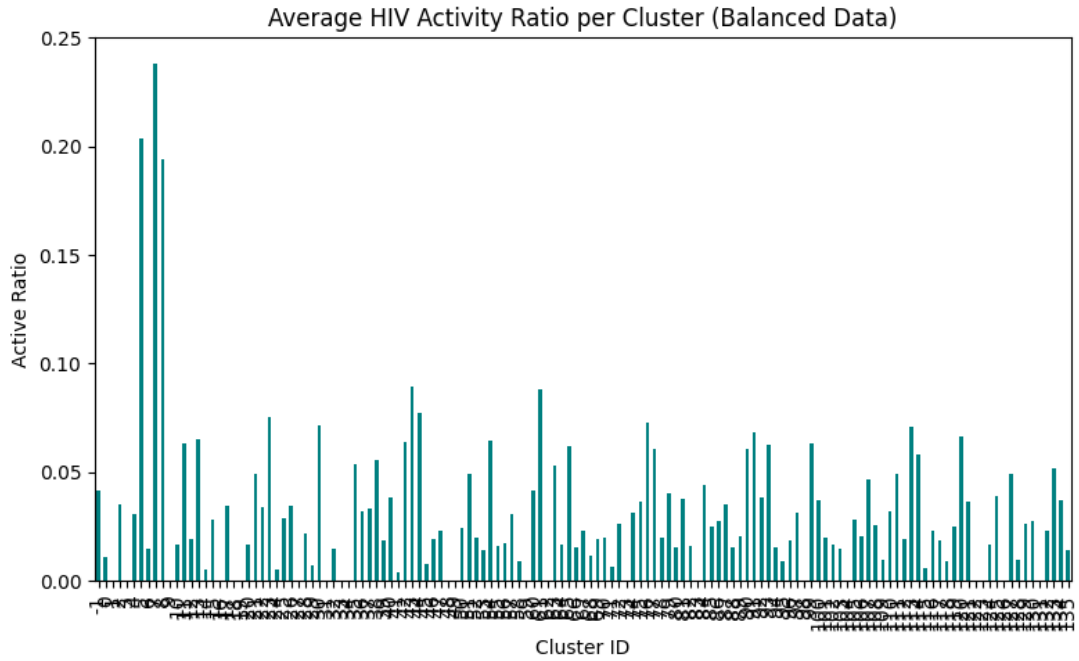
4.Results Analysis

4.1 Clusters Identification

Compared with traditional DBSCAN, HDBSCAN is more lightweight and memory-efficient[3]. It performs clustering through a hierarchical density tree rather than computing a full pairwise distance matrix, allowing it to handle large-scale, high-dimensional datasets without memory overflow. Thus, in the part we choose DBSCAN as main model.

HDBSCAN identified approximately 136 clusters (Cluster ID 0–135) along with around 19,000 noise points (Cluster -1), indicating the presence of many isolated or low-density samples — a characteristic consistent with the complex and non-uniform nature of chemical space.

Most clusters show low HIV activity ratios, while a few clusters (such as Cluster 5, 7, and 8) exhibit higher activity concentrations, suggesting that the model's latent space can partially distinguish between active and inactive molecular groups.



4.2 Functional Groups

The high-activity clusters (Cluster 5, 7, and 8) share several common functional features. Aromatic rings (phenyl) appear in over 90% of molecules across the three clusters, serving as common pharmacophoric structures that facilitate π–π interactions with target proteins[4]. Amide and phenol groups occur at around 60%, providing polarity and enhancing hydrogen bonding and solubility. Amine groups are relatively more abundant in Cluster 8 (about 0.51), suggesting that some active molecules may strengthen electrostatic interactions with receptors through basic sites.

4.3 Descriptor comparison (mean ± std)

For each descriptor, the mean ± standard deviation (std) was calculated, providing a quantitative summary of its central tendency and variability across clusters[8]. This comparison allows identifying structure–property trends among clusters — for example, whether high-activity clusters are associated with higher polarity, larger molecular size, or specific hydrogen-bonding capabilities.

The three clusters reveal two distinct structure–property trends: Clusters 5 and 7 represent a "high molecular weight, high polarity, strong-binding type," characterized by large frameworks and abundant polar functional groups that enable multiple hydrogen bonds and hydrophobic interactions with target sites. However, these

properties may limit membrane permeability and metabolic stability. In contrast, Cluster 8 exhibits a more compact structure with moderate molecular weight and polarity, showing a better balance between activity and drug-likeness. Based on the activity enrichment and scaffold analysis results, Cluster 8 is recommended as the priority candidate for further drug optimization.

Descriptor	Cluster 5 (Mean ± SD)	Cluster 7 (Mean ± SD)	Cluster 8 (Mean ± SD)
MolWt	898.804 ± 75.211	788.809 ± 73.633	639.792 ± 93.088
LogP	5.447 ± 2.787	5.172 ± 3.124	3.215 ± 3.469
TPSA	270.678 ± 53.703	244.287 ± 36.152	191.845 ± 33.030
HBD	6.296 ± 0.603	5.905 ± 0.754	5.051 ± 0.889
HBA	15.907 ± 1.137	13.571 ± 0.918	10.949 ± 1.196
RB	8.278 ± 3.547	8.238 ± 2.820	7.531 ± 2.668

5. Problems and Prospect

Most samples are concentrated in a narrow region, indicating that the current latent dimension (latent_size=2) may be too low, leading to excessive information compression. Increasing the latent size in future experiments may enhance the representational capacity and separability of the latent space.

Summary and reflection

The two experiments consistently indicate that highly active compounds commonly possess amide or aromatic scaffolds, conferring both planarity and molecular stability. They also exhibit strong polarity and multiple hydrogen bond donors and acceptors, enabling extensive hydrogen bonding and dipole interactions with polar residues of the target protein. Moreover, these molecules tend to have larger molecular weights and a certain degree of conformational flexibility, allowing efficient spatial accommodation

within the binding pocket and improved binding affinity.

Such structural characteristics are in line with the design principles of existing antiviral drugs, including HIV reverse transcriptase and protease inhibitors, which rely on amide linkages, aromatic systems, and polar interactions to achieve potent and stable binding[5]. Therefore, the shared features identified in this study not only highlight intrinsic patterns among highly active molecules but also provide valuable structural insights for designing new antiviral candidates.

Reference

[1] Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches[J]. Drug discovery today, 2007, 12(5-6): 225-233.

[2] Liu S, Cao S, Suarez M, et al. Multi-Level DBSCAN: A Hierarchical Density-Based Clustering Method for Analyzing Molecular Dynamics Simulation Trajectories[J]. BioRxiv, 2021: 2021.06. 09.447666.

[3] Stewart G, Al-Khassaweneh M. An implementation of the HDBSCAN* clustering algorithm[J]. Applied Sciences, 2022, 12(5): 2405.

[4] Li T T, Pannecouque C, De Clercq E, et al. Scaffold hopping in discovery of HIV-1 non-nucleoside reverse transcriptase inhibitors: from CH (CN)-DABOs to CH (CN)-DAPYs[J]. Molecules, 2020, 25(7): 1581.

[5] Takashiro E, Hayakawa I, Nitta T, et al. Structure–activity relationship of HIV-1 protease inhibitors containing α-hydroxy-β-amino acids. Detailed study of P1 site[J]. Bioorganic & medicinal chemistry, 1999, 7(9): 2063-2072.

[6] Rendon E, Alejo R, Castorena C, et al. Data sampling methods to deal with the big data multi-class imbalance problem[J]. Applied Sciences, 2020, 10(4): 1276.

[7] Hodson T O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not[J]. Geoscientific Model Development Discussions, 2022, 2022: 1-10.

[8] Wale N, Watson I A, Karypis G. Comparison of descriptor spaces for chemical compound retrieval and classification[J]. Knowledge and Information Systems, 2008, 14(3): 347-375.

Statements

This project utilized AI tools (ChatGPT, GPT-5 model; Claude) for technical assistance at certain stages, primarily in the following aspects:

1. Debugging and Environment Support:
Assisted in resolving code execution errors, path configuration issues, and environment compatibility problems.

2. Structural Optimization:
Helped refactor and encapsulate key components such as the Autoencoder and MolecularDescriptorClustering classes, improving code organization and clarity.

3. Analytical Suggestions:
Proposed extending the analysis beyond scaffold examination to include molecular descriptor statistics and functional group profiling.

4. Result Interpretation:
Provided support in the professional interpretation of molecular structures, chemical characteristics, and clustering outcomes.

5. The AI tool was used solely for technical and linguistic assistance.
All experimental design, data analysis, and final interpretations were conducted and verified independently by the researcher.