Molecular-substructure Deep Autoencoders Cluster Biomolecules into Novel Band-Shaped Substructure-Distinguished Bioactivity Clusters in 3D Latent Space

YING TAN
tan.ying@sz.tsinghua.edu.cn

Tsinghua University

Huazhang Ying

Xiang Wu

Chu Qin

Likun Zhang

Zhicheng Du

Jiaqi Liu

Yu Zong Chen

Tsinghua Shenzhen International Graduate School, Tsinghua University https://orcid.org/0000-0002-5473-8022

Article

Keywords: Bioactive molecule, DAEs, substructure

Posted Date: June 16th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6755378/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Unsupervised deep autoencoders (DAEs) are useful for data clustering and visualization. DAE-derived data clusters are typically visualized by dimensionality reduction methods, which have some degree of visual distortions that pose difficulties in revealing intrinsic cluster patterns. Here, we developed substructure-based molecular-fingerprint DAEs (MolF-DAEs) to cluster 1.9 million bioactive molecules (biomolecules) in 3D latent space (3DLSpace), where data clusters can be straightforwardly visualized. MolF-DAEs developed with three established sets of molecular fingerprints consistently cluster biomolecules with 96.1–97.6% reconstruction rate. In 3DLSpace, the biomolecules cluster into novel substructure-distinguished bioactivity-relevant band-shaped clusters. Each cluster is dominated by the biomolecules of specific substructure combinations. These in-cluster biomolecules are of varying molecular structures but frequently form a limited number of bioactivity classes. Our study suggests that unsupervised deep clustering in 3DLSpace is useful for visually revealing the intrinsic data distribution patterns and functionally relevant data clusters.

Introduction

Unsupervised learning such as deep clustering methods has been widely applied in real-life statistical analysis such as pattern recognition[1], image processing[2], and knowledge discovery tasks such as bioactive molecule (biomolecules) clustering[3,4], genomics data mining[5], and disease diagnosis[6]. One application of deep clustering is in drug discovery, where effective clustering of biomolecules with respect to common molecular determinants facilitates the mapping of pharmacological chemical space[7] and the investigation of structure-activity relationships[8].

Various clustering methods have been developed based on the fundamental data features or their linearly/non-linearly transformed variants[9], such as K-means[10,11], hierarchical clustering[12] and spectral clustering. Moreover, deep autoencoders (DAEs) are highly useful for deep and complex clustering tasks[13,14]. Due to the high dimensionality, sparsity and variance of data features, these clustering methods rely on feature representation and dimensionality reduction techniques[14,15]. Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and t-Distributed Stochastic Neighbor Embedding (t-SNE)[16] are the most common algorithms used as a preprocessing step to provide useful cluster patterns. Under these methods, the visualization of data clusters and the subsequent analysis may be affected by visual distortion in the low-dimensional space. Minor visual distortions may in some cases affect the quality of cluster analysis. For example, minor differences in biomolecular structures (i.e. minor differences in the separation of the cluster neighbors) may lead to substantial changes in bioactivity targets and bioactivity values [17]. There is a need for effective methods to both cluster data and visualize the undistorted cluster patterns.

DAEs [13,14] may be potentially explored for data clustering and undistorted visualization. It captures nonlinear relationships of complex patterns while preserving both local and global characteristics[18,19]. In order for undistorted visualization of the DAE-derived data clusters, one may consider the construction of DAEs with 3DLSpace, where the data clusters can be straightforwardly visualized in the 3DLSpace without data distortion. A question is whether DAEs can meaningfully cluster data in 3DLSpace. Here we developed molecular fingerprint deep autoencoders (MolF-DAEs) to demonstrate the clustering ability of DAEs in 3DLSpace. We further revealed the DAE-derived cluster patterns of bioactive molecules and discussed their potential implications to drug discovery tasks.

MolF-DAEs consist of symmetric fully connected encoders and decoders, which were trained by 1.9 million biomolecules from the ChEMBL database[20] represented by three sets of molecular fingerprints (MFs). The three sets of DAEs are the PubChem molecular fingerprint model (PubChemFPM), MACCS keys fingerprint model (MACCSFPM), and 2D pharmacophore fingerprints model (PharmacoPFPM)[21] (Fig. 1). Compared with existing methods, MolF-DAEs do not require additional dimensionality reduction methods. Additionally, we developed a chemical space navigation simulation software Chempack for displaying and analyzing the band cluster landscapes. Rather than solving a specific downstream task prediction, MolF-DAEs aim at mining reliable experimentally obtained high activity data to evaluate activity-related compound spatial and organism target spatial. This method can be migrated to other types of sparse high-dimensional data mining.

Results

1. High Accuracy of Deep Autoencoders in Reconstructing Three Fingerprint Feature Maps

This work demonstrates the high efficacy of deep autoencoders in reconstructing biomolecule fingerprint feature maps (Fmaps). In order to increase the efficiency, we focus on the molecules presenting preliminary biological activity. In total, a 1.9 million dataset with high bioactivity (Potency Values ≤10 μM) was collected in the ChEMBL database. Three deep autoencoders are constructed to encode three sets of molecular fingerprint features, including PubChemFP and MACCSFP based on SMILES arbitrary target specification (SMARTS), and PharmacoPFP based on pharmacophore. The reconstruction rates for PubChemFPM, MACCSFPM, and PharmacoPFPM reach 97.55%, 96.10% and 97.55% respectively, surpassing the reported reconstruction rates of 95.3%-96.4% from a latent space dimension of 196 in a variational autoencoder (VAE) trained on 250,000 drug-like molecules[4]. It indicates the precision of models in constructing two-dimensional Fmaps, thereby establishing their reliability in interpreting the molecule physicochemical properties, structural fragments, and spatial distributions. Taking the PubChemFP model as an example, the loss values stabilize around 0.0245 after 100 training epochs (Fig.2a). Despite the correlation between the parameter count and the performance of the models, the minimum MSE may not necessarily require the maximum number of parameters (Fig.2b). Compared with SMILES, the Fmaps are of higher distinguishment among molecules. As the MSE decreases, the visual improvement in the reconstruction of Fmaps becomes apparent (Fig.2c, Supplementary Fig.1-6). Despite the low dimensional vector cause difficulties in accurately reconstructing fingerprints[22], the majority of the original data characteristics preserved postreconstruction, barring slight deviations in some local features. This attributed to the extensive training data and feature representation methodologies.

2. Three sets of molecular fingerprints FP-DAEs exhibit band-shaped clustering patterns in 3DLSpace

It is intriguing to see the distribution landscape in latent space based on substructures and pharmacophores. Overall, these distributions are orderly and highly consistent with human knowledge. MolF-DAEs achieve optimal distribution in 3DLSpace as training progresses (Fig.3a-c). It exhibits a discontinuous spatial distribution with distinct boundaries. Internally, molecules are arranged into bands and each of them originates from a common central region. This arrangements was observed in nonlinear algorithms such as UMAP, representing mappings dominated by the strong effects of major gradient features[23,24]. It is observed for the first time in DAEs.

The latent space exhibits distinct, island-like regions—dense clusters where specific molecular substructures or pharmacophore labels are uniquely enriched. These regions emerge as training progresses, separating from the general molecular distribution to form cohesive zones, each representing variations of a common molecular feature. Target type is a molecular feature characterization received widespread attention[25]. It's reports that over 50% of drug design targets are concentrated in four categories, including kinase, protease, nuclear receptor and G-protein coupled receptor (GPCR) [26]. However, these target families cover only 1.45%-6.42% of experimental verified biomolecules in the 1.9 million compound database. This indicates that the space of biomolecules remains vast. In MolF-DAEs, without any prior target information, these targets naturally appear in relatively isolated regions (Fig.3d-g). For example, in some of these regions, the concentrations of specific targets reached the following values: kinase (71%), protease (71%), and GPCR (54%). Therefore, MolF-DAEs demonstrate exceptional effectiveness in identifying target-specific clusters. Within the PubChemFPM, kinase inhibitors are concentrated in the upper region in long islands, while protease inhibitors are concentrated in the lower region. GPCR has fewer known drugs and more diverse natural ligands[27]. The clustering of GPCR appears dispersed and concentrated in short islands (Fig.3d, Supplementary Fig.7).

The three most common literature-reported molecular fingerprint features are used (Supplementary Figs. 7-9). Each of them exhibits unique clustering patterns. PubChemFPM captures the substructural features of molecules, with target-specific bands with the clearest separation. In contrast, MACCSFPM mainly focuses on the overall structure of molecules, closely connected among clusters. PharmacoPFPM mainly describes the position, spatial relationship, and interaction pattern of pharmacophores in molecules. It exhibits many distinctly isolated short islands in addition to the band structures. The performance of downstream tasks is related to factors like fingerprint types, dimensions, compression, and redundancy[22].

3. Bands exhibit significant differences in the potential biotherapeutic targets within the 3DLSpace

The in-cluster biomolecules are of varying molecular structures but frequently form a limited number of bioactivity classes. Significant concentrations are captured in the scale of substructures, physical properties, and targeting families. For subsequent analyses and applications, we manually select six representative bands with clear boundaries spanning the compound space, across all fingerprint channels (band 1-4 PubChemFPM, band 5: MACCSFM, band 6: PharmacoPFPM) and privileged target islands (band 1-2, 5-6 kinase, band 3 protease, band 4 GPCR). Although there is a fair number of confusing ligands (gray), the four major target classes account for up to 34% of points (Fig.4a and Supplementary Data 8). Kinase is a class of targets with relatively conserved binding pockets, causing widespread off-target effects. Kinase inhibitor takes 69.22%-84.01% in bands 1-2, 5-6 (Fig.4a) and significant changes appear among families (Fig.4b) and groups (Fig.4c). In four kinase enriched bands, members of the Janus kinase (JakA) and Receptor tyrosine kinase (RTK) group, such as the Epidermal growth factor receptor (EGFR) and fibroblast growth factor receptor (FGFR) family exhibit highest proportions. The Tec family is uniquely enriched in PubChemFPM band 1. EGFR inhibitors such as Erlotinib and Gefitinib are developed as anticancer therapeutics, mainly consisting of rings as a basic skeleton[28,29]. They are enriched within the kinase cluster and positioned closely to each other, with a distance of 25.42. The top 10 families enriched in PharmacoPFPM are more concentrated and closer in position on inter-group and evolutionary trees (Supplementary Fig.11). Evolutionarily closely related kinase families tend to have similar core structures, likely to cause cross-effects. Conversely, the PubChemFP band covers a broader range of families, including a new enrichment inhibitor group with fewer known inhibitor studies such as Tyrosine kinase-like (TKL), Calcium/calmodulin-dependent protein kinase (CAMK) and the remainder[27]. The degree of enrichment varies significantly in remainders. PharmacoPFP band 5 enriches 30.61% of the Calcium/calmodulin-dependent protein kinase group (CAMK) with fewer reported researches[30,31]. It indicates that relevant structures with similar activities are sensitively captured by the model in undistorted clusters. This contributes to novel target inhibitor pattern researches.

4. Core substructures combination explains undistorted band-shape cluster organization

There is core structural unity and local residue diversity in 10^2 sample size, suggesting the high-quality undistorted distribution pattern. Privileged target types exist, while the principle behind target label clustering is limited to FMap-related substructural classes. GPCR binding fragments share a conserved structural scaffold, whereas kinase inhibitors exhibit more variations in the scaffold and substituents. However, distinct GPCR islands with clear boundaries and structural similarities were identified in kinase-enriched band 1 (Fig.4d). 95.21% of molecules in band 1 share hydrogen bond sites N-C-N-C including the GPCR islands, while other GPCR-enriched bands seldom take this substructure (3.23%). In another kinase-enriched band, the N-C-N-C-C-N structure is over twice as prevalent compared to band 1. Molecules in the kinase island typically exhibit higher LogP, while both molecular weight and LogP of biomolecules in the GPCR islands are lower. Thus, MolF-DAEs separate bands based on various substructural or pharmacophore features. This further explains the intrinsic reasons for cluster-target relation. Furthermore, this data-driven clustering contributes to drug novelty evaluation. Molecules contain amide bonds (-CONH-) with hydrogen bonding, and methoxy-substituted phenyl rings are enriched in kinase island 1. The structure is common in various drugs, including anticancer, antibacterial, and anti-inflammatory drugs like Amitriptyline. Methoxy-substituted phenyl rings are relatively easy to introduce in organic synthesis, and methoxy groups increase lipophilicity, affecting drug metabolism and absorption. In kinase island 2, more substituents like methoxy (-OCH3), fluorine (-F), and chlorine (-Cl) are present. Notably, the gray biomolecules are highly likely to be new kinase biomolecules.

To explain the observed undistorted clustering, we summarize the common functional structural modes. It provides a basis for binding studies. Structural analysis reveals a widespread substructural pattern. Each band cluster primarily consists of molecules with unique scaffold or substructure combinations, with minimal overlap with other clusters. The kinase-specific substructure mode involves a combination of two elements: core hydrophobic ring elements (blue) and core hydrogen site elements (red) (Fig.5a). These substructures account for less than 15% of the compounds in other target bands (Fig.5c, Supplementary Fig.12b). The linear structures serve as hydrogen donors and acceptors. It is decisive for the overall binding strength and positioning. Two kinds of linear structure are observed, one is the linear framework and its variants, and the other is the Y-shaped frameworks and their variants. These elements like Y-shaped N-O-N, N-S-N, N-O-CN, or L-shaped NC-N-CN, along with core hydrophobic ring elements, are observed in Lapatinib, Imatinib and Sorafenib. Hydrophobic ring elements contain 6-membered or 5-membered carbon rings, which is important for hydrophobicity and aids in overall stabilization. The ring elements are connected by one or more carbon chains to the core elements. The question is which molecular features are crucial for selective kinase targeting and potency. Compared with the overall common background, we counted the frequency of occurrence of these substructures in 1.9 million molecules and that on the band, respectively. In kinase- specific inhibitors bands, the core element combination reaches up to 22.01 fold to background (Fig.5b, supplementary Fig. 9). This indicates that MolF-DAEs highly select this bioactive substructure as an important feature. Overall, in PubChemFPM regions, the linear carbon-nitrogen substructure (NC-N-C) appears in almost all known kinase inhibitors (band 1: 100%, band 2: 99.8%) and other (band1: 95.2%, band2: 95.1%). And most of the kinase inhibitors are connected to at least one six-membered ring (band 1: 62.2%, band 2: 40.4%) or five-membered ring (band 1: 43.4%, band 2: 37.1%). MACCSFPM covers fewer chemical substructures compared to the PubChemFPM, indicating a more concentrated distribution. There is no Y-shaped substructure connected to six-membered rings in band 5. Pharmacophore fingerprint primarily focuses on describing the position, spatial relationships, and interaction patterns of pharmacophores in molecules. The clustering of PharmacoPFPM in 3DLSpace is more pronounced, with relatively high proportions in various feature substructures.

Protease is another target type significantly independent in the band. Peptide chains are prevalent, exhibiting variations in C-chain length and element substitutions. In fact, protease inhibitors contest with natural substrates peptides while not being degraded by proteases. Some unidentified ligands with long peptide chains are captured in the band, such as CHEMBL100202 and CHEMBL102898. In terms of property labeling, this is at variance with the distribution of overall physicochemical properties of approved drugs reported[32]. Similar to peptide drug, protease-privilege bands generally have high molecular weights, lower logarithmic partition coefficients (LogP) values (indicating greater hydrophilicity), and higher topological polar surface area (TPSA). For example, 27% of the molecules exhibiting a LogP below 0, which is significantly higher than average for approved drugs (focus on 1-3). The amide bonding increases the polarity of the molecule and the number of hydrogen bonding donors/acceptors. These differences in structure and physicochemical properties tend to have lower quantitative estimates of drug-likeness (QED) values , averaging around 0.1—well below the reported average of 0.35. This underscores the unique structural and drug-forming properties of this cluster[32].However, these features, while conducive to protease activity, pose additional challenges in pharmacokinetics for molecule in this band, similar to those faced by peptide drugs.

There are GPCR concentrated blocks in band 1 and band 4 despite the fact that the number of compounds is less than half that of kinase and protease. In contrast to the kinase conserved pocket, the natural substrates of GPCR are more mixed, including both nucleic acid substrates and peptide substrates. In 3DLSpace, GPCR-specific clusters are mostly found in the form of short islands, with more dispersed molecular clustering. Some are distributed within other target clusters, implying that they have similar functional groups or competing substrates. However, a small number of independent long island GPCR bands still exist in band1. Structures within islands are often very similar in specific lengths of core hydrogen bonding elements (Y-shaped NCO, NCN, NCCO), combined with core hydrophobic ring elements (R and RN)[33]. Islands differ mainly in substituents, such as O/N-rich islands 1 and 2, and F/Cl-rich island 2. GPCRs have binding pockets that vary significantly in nature, with F providing strong liposolubility possibly related to binding to the water transport pocket in the transmembrane region, and O being closely associated with the extracellular region.

5. Relevance and distinction of sub-structural features with respect to the literature-reported privileged pharmacophores and drug-binding mode

A question is raised about the relevance of the DAE-captured sub-structural features of an individual band cluster to the selected bioactivities of the band cluster. Through literature-reported kinase structures, the enriched categories are highly consistent with those reported[34]. Some substructural features of the band clusters comprise key frameworks in literature-reported kinase-binding modes of kinase inhibitor drugs, pharmacophores of kinase frequent hitters, and privileged fragments of kinase inhibitors. The L- and Y-shaped regions are usually binding to a stabilized presence of a hinge area. Core hydrogen bonding structures in different lengths allow for the choice of four sites that bind 2 to 3 sites to the hinge region or gatekeepers. Hydrophobic rings, which are larger and segmented into two parts, are more likely to extend into the E0 or back pocket to form stacking interactions. Different regions with unique substructure patterns are selective in their binding conformation. Structural differences in the remaining parts allow binding pockets of different conformations[35]. The reported kinase-binding conformations of band 2 with short/Y-shape hydrogen sites (PDB: 9JI, 3JW, GD9, P06, 6S1, and RXT) are in the front pocket, without the back pocket. Complexes with multiple long-chain hydrogen sites in band 1 (PDB: TZ0, R1L and UCW) occupy both the front and back pockets (Fig.6).

The sub-structural characterization provided by MolF-DAEs is closely consistent with the framework of selected bioactivities and the concentration of these biomolecules within the bands. This correlation is notably in line with the finding that kinase-specific fragments can enhance kinase inhibitors by 5-fold[36], with sub-structures within individual bands exhibiting enrichments exceeding 25-fold. This concurrence also echoes reports indicating that certain specific molecular scaffolds can exhibit activity against multiple target classes. However, the substructural elements captured by MolF-DAEs diverge from the pharmacogenetic and privileged fragments outlined in the literature in one aspect. While MolF-DAEs capture the fundamental substructural elements and their structural variations, which collectively define the framework of pharmacogenetic or privileged fragments, kinase-specific structures documented in literature often manifest as specific combinations of fragments, such as bis-aryl-NH-linked fragments and biphenyl ether scaffolds[37]. Consequently, by assimilating the foundational elements of structural frameworks, MolF-DAEs possess the capacity to capture a broader spectrum of pharmacogenetically and conventionally framed specific structures, thereby clustering them into individual band clusters.

Discussion

DAE is a potential deep learning-based strategy to solve the undistorted presentation and clustering. DAEs have a predictive performance on high-dimensional datasets[38] and successfully tackled challenging tasks on millions of training samples[14,39,40]. In contrast to a fixed kernel function in nonlinear functions, autoencoders are learned by optimizing the reconstruction error[41]. Reconstruction effects reflect the ability to represent potential space.

For data with high complexity such as drugs, it seems difficult for DAE to directly downscale to 3D space because information loss is inevitable. Thus current data such as MNIST can only be downscaled from 28×28×1 images to 128. Effort is focused on the joint methods in DAE that have been developed to handle the space distortion challenge[42,43]. Visualization is satisfied by additional downscaling methods. Deep autoencoder (DAE) is an undistorted clustering method to handle high dimensional datasets without additional clustering strategy. In contrast, it performs badly based on the 1.9 million biomolecule dataset when combining traditional clustering methods (UMAP and PCA) to 3-dimensional space from 128-dimensional DAEs latent spaces. Molecules exhibit a typical spherical distribution in 3D space. It doesn’t perform well and escapes the same target and substructure separation characteristics as MolF-DAE (Supplementary Fig. 21).

In biological experiments, the research of biomolecules focuses on dozens or hundreds of data. Supervised learning research of biomolecules focuses on tens of thousands of high-quality target or disease data. Unsupervised learning breaks through the limitations of data quality. This work demonstrates significant improvements in both the size of the training dataset and the systematic utilization of physicochemical property feature dimensions. Surprisingly, all three sets of molecular fingerprints for various target types exhibit a characteristic radial distribution emanating from the origin. They exhibit target-specific clusters that in turn intrinsically reflect a more essential classification based on molecular structure. This has potential implications for biomolecule classification and research. The information obtained from the model substructures is highly consistent with human knowledge, enabling the possibility of subdividing the biomolecules into refined subclasses. It offers crucial clues to understand the relationship between the structure and activity of drugs. Meanwhile, MolF-DAEs offer the possibility of exploring the chemical space more impartially by effectively acquiring meaningful latent space, free from human rationality or bias. Finally, this work provides interpretability to the clustering distribution of unsupervised models, while also aiding in tasks such as understanding target-related drug structures, identifying potential drug candidates, and facilitating drug repurposing efforts.

MolF-DAEs offer several possible directions of application. From a drug perspective, the conserved structures of a specific collection of molecules are used as a method for ligand-based virtual screening to discover new drugs for specific targets. Evaluate the novelty of bioactive skeleton and diversity of substituents. To guide the generation of new structures. From the target point of view, the biomolecule structure and drug diversity of the target can be evaluated, and the potential off-target targets of the compound list can be predicted. Evaluation of drug cross-reactivity of multiple proteins, etc.

In the field of data analysis, compounds are a class of data types that are rich in structural information and have high-dimensional feature representations. In the future, based on this data can be migrated to more knowledge-related types of sparse high-dimensional data for non-destructive spatial clustering display and data mining.

Method

Data collection and labeling. Medium to high biomolecules with IC50, EC50 and Ki ≤ 10 μM through experimental methods such as MTT assay, kinase activity test, etc., are selected as datasets, from the pharmacochemical database ChEMBL[44,45]. It covers compounds from the preclinical to the approved stage. There are 1,943,048 biomolecules in total.

Label standing for the verified target is used for visualization in 3DLSpace. Molecules with activity against four major drug target classes were queried, including kinase, protease, GPCR, nuclear receptor, etc. 124,632 biomolecules targeting kinase (red), 92,040 targeting protease (green), 33,718 targeting GPCR (Cyan) and 28,216 targeting nuclear receptor (purple) are obtained. Compounds with multiple target label values were excluded during 3D visualization. 1,658,503 biomolecules targeting other types of targets.

Construction of molecular fingerprints Fmaps. We selected three molecular fingerprints with the highest citations in 1410 literature by 2024, according to the PubMed database, two substructure-key SMARTS-based features (PubChemFP, 192 bits and MACCSFP, 476 bits), and pharmacophore-based features (PharmacoPFP, 298 bits)[46]. These fingerprints are generated based on MolMap, which is a new method of molecular feature generation based on manifold learning[21,47]. We use PyBioMed to remove part of the

PubChem molecular fingerprint unique thermal code infrequently, reducing the original 881 dimensions to 733 dimensions[48].

Optimization of the parameters of MolF-DAEs. A pair of complementary DNNs are adopted, with an encoder as an extractor to convert Fmaps into 3DLSpace for clustering the molecules, and a decoder to convert the latent codes back to the original FMaps for optimizing the autoencoder. Adam optimizer is adopted, with the loss function initially using binary cross entropy and MSE successively for 100 epochs training expectedly. The trend of the loss function (binary cross entropy (left) and MSE (right) process of loss function.

$$L_{rec} = \min \frac{1}{n} \sum {}_{i=1}^{n} \| x_i - \varphi_{r} \left(\varphi_{e} \left(x_i \right) \right) \|^2$$

Where $\varphi_{e}(.)$ and $\varphi_{r}(.)$represents the encoder network and decoder network of MolF-DAE respectively.

The hyperparameters of the autoencoder were optimized by the tree-structured estimator approach in two phases[49]. In phase 1, the number of layers varied from 2 to 100, the number of nodes in the first hidden layer was set at 2048 followed by up to 50% reductions in the subsequent layers. It proceeds until the reconstruction rate reaches > 95% (percent of position-to-position reconstruction of the original molecular fingerprints), which is comparable to the reported 95.3%-96.4% reconstruction rate of an autoencoder trained on 1,937,109 drug-like molecules[4]. In phase 2, the number of nodes in the hidden layers was fine-tuned. Phase 2 proceeds until the reconstruction rate reaches optimal value. The optimal parameters are detailed in Supplementary Table 10. In PuChemFPM, the encoder consists of a 5-layer fully connected neural network (DNN). In Encoder, each dense layer is followed by a ReLU activation function. The final layer connects to a fully connected layer with 3 nodes. Each dense layer in the decoder is followed by a ReLU activation function, except for the last dense layer, which employs a Sigmoid activation function. The total parameter count is 2,697,564, approximately 2.7 million.

Comparing the sizes of individual images, MACCSFP constitutes only a quarter of PubChemFP. Consequently, models with original parameters possess excessive total parameters for the new dataset, leading to suboptimal Fmaps reconstruction. Models with smaller total parameters were adopted, and the parameter space was gradually expanded to find the optimal parameters. During optimization, a total parameter count of 2 million yields better training results. The composition of the 2 million parameters network was continuously adjusted. It’s an unstable model in the task of reconstructing images from large datasets, although the variation range of the stabilized MSE is small and within an acceptable range. Therefore, the choice between these parameters has minimal impact on the final model selection. The influence of model network depth, encoder structure, and total parameter count on the reconstruction effect outweighs the impact of the model's inherent instability.

Chempack Software Demonstration

Accessing spatial data linearly presents challenges in quickly addressing spatial queries during the data retrieval process and in conducting statistical analyses based on the macroscopic distribution of data in space. To address these issues, we have developed Chempack software for fast navigation and simultaneous display of the DAE-generated distribution landscapes of up to 50 million molecules in the 3-dimensional latent chemical space. The molecules within an on-screen subspace (a movable cubic box) are displayed as bright spheres (in default grey color) embedded in the black background space. Subsets of molecules may be highlighted by user-specified colors (via color selector). The molecules outside the moveable cubic box are un-displayed unless the box is moved into the local subspace. A multi-layer iterative data retrieval and display algorithm was employed for displaying the spheres within a local cubic box of 1/ 4N-1 (N = 1–8) of the volume of the global cubic box defined by the input spheres. The following procedure created the global and local cubic box. A user manual for Chempack is provided in the supplementary materials.

In this paper, we utilize Chempack software to visualize 3DLSpace by inputting three numerical values as coordinate values. With the addition of labels previously annotated for 1.9 million chemical biomolecules, these serve as the original data coordinates, enabling visualization in 3DLSpace. Different labels are represented by different colors for visual differentiation. Chempack software assigns distinct colors to different target types, where each target type corresponds to a specific color label: kinase (red), protease (green), nuclear receptor (purple), G protein-coupled receptor (GPCR) (cyan), and other targets (gray).

Declarations

Acknowledgments

The authors acknowledge the financial support from the National Key Research and Development Program of China (grant number: 2023YFA0913600) and Research Fund of Guangdong Province (grant number: 2024A1515011906).

Code availability

All code used in this paper is open source. Code for MolF-DAEs is available via GitHub at https://github.com/HuazhangYing/MolF-DAEs. Code for feature representation is available via GitHub at https://github.com/shenwanxiang/bidd-molmap. The dataset for 1.9 million bioactivate molecules is available at https://drive.google.com/drive/folders/1oPpz3biogeAEoa9-RydxdLDBcZQd-6o6?usp=drive_link. Resources for the 3D visualization software Chempack are openly accessible. For access or further information, please contact the authors.

Author contributions

Y.T., Y.C., and C.Q. conceived the project. H.Y. and X.W. develop the methodology, and performed experiments and analysis. H.Y. wrote the paper. H.Y., Y.C. and Y.T. contributed to the revision of the manuscript.

Competing Interests statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Oyedotun, O. & Dimililer, K. Pattern Recognition: Invariance Learning in Convolutional Auto Encoder Network. *International Journal of Image, Graphics and Signal Processing* 8, 19-27 (2016). https://doi.org:10.5815/ijigsp.2016.03.03
2. Dundar, A., Jin, J. & Culurciello, E. Convolutional Clustering for Unsupervised Learning. *ArXiv* abs/1511.06241 (2015).
3. Polanski, J. Unsupervised Learning in Drug Design from Self-Organization to Deep Chemistry. *International Journal of Molecular Sciences* 23, 2797 (2022).
4. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* 4, 268-276 (2018). https://doi.org:10.1021/acscentsci.7b00572
5. Clarke, R. *et al.* The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer* 8, 37-49 (2008). https://doi.org:10.1038/nrc2294
6. Chen, L., Saykin, A. J., Yao, B. & Zhao, F. Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood. *Computational and Structural Biotechnology Journal* 20, 5761-5774 (2022). https://doi.org:https://doi.org/10.1016/j.csbj.2022.10.016
7. Mullowney, M. W. *et al.* Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery* 22, 895-916 (2023). https://doi.org:10.1038/s41573-023-00774-7
8. Tropsha, A., Isayev, O., Varnek, A., Schneider, G. & Cherkasov, A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nature Reviews Drug Discovery* 23, 141-155 (2024). https://doi.org:10.1038/s41573-023-00832-0
9. Song, C., Liu, F., Huang, Y., Wang, L. & Tan, T. in *Proceedings, Part I, of the 18th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - Volume 8258* 117–124 (Springer-Verlag, Havana, Cuba, 2013).
10. Walters, S. J. & Campbell, M. J. The use of bootstrap methods for analysing Health-Related Quality of Life outcomes (particularly the SF-36). *Health Qual Life Outcomes* 2, 70 (2004). https://doi.org:10.1186/1477-7525-2-70
11. Eckhardt, C. M. *et al.* Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surg Sports Traumatol Arthrosc* 31, 376-381 (2023). https://doi.org:10.1007/s00167-022-07233-7

12. Altman, N. & Krzywinski, M. Clustering. *Nature Methods* 14, 545-546 (2017). https://doi.org:10.1038/nmeth.4299

13. Kamal, I. M. & Bae, H. Super-encoder with cooperative autoencoder networks. *Pattern Recognition* 126, 108562 (2022). https://doi.org:https://doi.org/10.1016/j.patcog.2022.108562

14. Hinton, G. E. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 504-507 (2006). https://doi.org:doi:10.1126/science.1127647

15. Steinbach, M., Ertöz, L. & Kumar, V. in *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition* (ed Luc T. Wille) 273-309 (Springer Berlin Heidelberg, 2004).

16. Yu, W., Wang, R., Nie, F. & Wang, F. Multi-view embedded clustering with unsupervised trace ratio LDA. *Neurocomputing* 315, 169-176 (2018). https://doi.org:https://doi.org/10.1016/j.neucom.2018.07.014

17. Stumpfe, D. & Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* 55, 2932-2942 (2012). https://doi.org:10.1021/jm201706b

18. Han, Z. *et al.* Mesh Convolutional Restricted Boltzmann Machines for Unsupervised Learning of Features With Structure Preservation on 3-D Meshes. *IEEE Trans Neural Netw Learn Syst* 28, 2268-2281 (2017). https://doi.org:10.1109/tnnls.2016.2582532

19. Guo, X., Liu, X., Zhu, E. & Yin, J. 373-382 (Springer International Publishing).

20. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* 52, D1180-D1192 (2023). https://doi.org:10.1093/nar/gkad1004

21. Shen, W. X. *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nature Machine Intelligence* 3, 334-343 (2021). https://doi.org:10.1038/s42256-021-00301-6

22. Ilnicka, A. & Schneider, G. Compression of molecular fingerprints with autoencoder networks. *Molecular Informatics* 42, 2300059 (2023). https://doi.org:https://doi.org/10.1002/minf.202300059

23. Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* 37, 1482-1492 (2019). https://doi.org:10.1038/s41587-019-0336-3

24. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* abs/1802.03426 (2018).

25. Li, Y. H. *et al.* Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Briefings in Bioinformatics* 21, 649-662 (2019). https://doi.org:10.1093/bib/bby130

26. Vasaikar, S., Bhatia, P., Bhatia, P. & Yaiw, K.-C. Complementary Approaches to Existing Target Based Drug Discovery for Identifying Novel Drug Targets. *Biomedicines* 4, 27 (2016). https://doi.org:10.3390/biomedicines4040027

27. Attwood, M. M., Fabbro, D., Sokolov, A. V., Knapp, S. & Schiöth, H. B. Trends in kinase drug discovery: targets, indications and inhibitor design. *Nature Reviews Drug Discovery* 20, 839-861 (2021).

https://doi.org:10.1038/s41573-021-00252-y

28. Kumar, V. *et al.* Role of Tyrosine Kinases and their Inhibitors in Cancer Therapy: A Comprehensive Review. *Curr Med Chem* 30, 1464-1481 (2023). https://doi.org:10.2174/0929867329666220727122952
29. Shaban, N., Kamashev, D., Emelianova, A. & Buzdin, A. Targeted Inhibitors of EGFR: Structure, Biology, Biomarkers, and Clinical Applications. *Cells* 13 (2023). https://doi.org:10.3390/cells13010047
30. Santiago, A. D. S. *et al.* Structural Analysis of Inhibitor Binding to CAMKK1 Identifies Features Necessary for Design of Specific Inhibitors. *Sci Rep* 8, 14800 (2018). https://doi.org:10.1038/s41598-018-33043-4
31. Profeta, G. S. *et al.* Binding and structural analyses of potent inhibitors of the human Ca(2+)/calmodulin dependent protein kinase kinase 2 (CAMKK2) identified from a collection of commercially-available kinase inhibitors. *Sci Rep* 9, 16452 (2019). https://doi.org:10.1038/s41598-019-52795-1
32. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry* 4, 90-98 (2012). https://doi.org:10.1038/nchem.1243
33. Bondensgaard, K. *et al.* Recognition of Privileged Structures by G-Protein Coupled Receptors. *Journal of Medicinal Chemistry* 47, 888-899 (2004). https://doi.org:10.1021/jm0309452
34. Liao, J. J.-L. Molecular Recognition of Protein Kinase Binding Pockets for Design of Potent and Selective Kinase Inhibitors. *Journal of Medicinal Chemistry* 50, 409-424 (2007). https://doi.org:10.1021/jm0608107
35. van Linden, O. P. J., Kooistra, A. J., Leurs, R., de Esch, I. J. P. & de Graaf, C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space. *Journal of Medicinal Chemistry* 57, 249-277 (2014). https://doi.org:10.1021/jm400378w
36. Aronov, A. M., McClain, B., Moody, C. S. & Murcko, M. A. Kinase-likeness and Kinase-Privileged Fragments: Toward Virtual Polypharmacology. *Journal of Medicinal Chemistry* 51, 1214-1222 (2008). https://doi.org:10.1021/jm701021b
37. Schneider, P. & Schneider, G. Privileged Structures Revisited. *Angewandte Chemie International Edition* 56, 7971-7974 (2017). https://doi.org:https://doi.org/10.1002/anie.201702816
38. Ilnicka, A. & Schneider, G. Designing molecules with autoencoder networks. *Nature Computational Science* 3, 922-933 (2023). https://doi.org:10.1038/s43588-023-00548-6
39. Gebru, T. *et al.* Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences* 114, 13108-13113 (2017). https://doi.org:doi:10.1073/pnas.1700035114
40. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* 550, 354-359 (2017). https://doi.org:10.1038/nature24270
41. Lu, S. & Li, R. DAC: Deep Autoencoder-based Clustering, a General Deep Learning Framework of Representation Learning. *ArXiv* abs/2102.07472 (2021).

42. Ren, Y. *et al.* Deep Clustering: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21 (2024). https://doi.org:10.1109/TNNLS.2024.3403155

43. Xie, J., Girshick, R. & Farhadi, A. in *Proceedings of The 33rd International Conference on Machine Learning* Vol. 48 (eds Balcan Maria Florina & Q. Weinberger Kilian) 478--487 (PMLR, Proceedings of Machine Learning Research, 2016).

44. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 47, D930-d940 (2019). https://doi.org:10.1093/nar/gky1075

45. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* 45, D945-d954 (2017). https://doi.org:10.1093/nar/gkw1074

46. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* 59, 3370-3388 (2019). https://doi.org:10.1021/acs.jcim.9b00237

47. Shen, W. X. *et al.* AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks. *Nucleic Acids Research* 50, e45-e45 (2022). https://doi.org:10.1093/nar/gkac010

48. Dong, J. *et al.* PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *J Cheminform* 10, 16 (2018). https://doi.org:10.1186/s13321-018-0270-2

49. *Proceedings of the 24th International Conference on Neural Information Processing Systems*. (Curran Associates Inc., 2011).

50. Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. P. & Kooistra, A. J. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Research* 49, D562-D569 (2020). https://doi.org:10.1093/nar/gkaa895

Figures

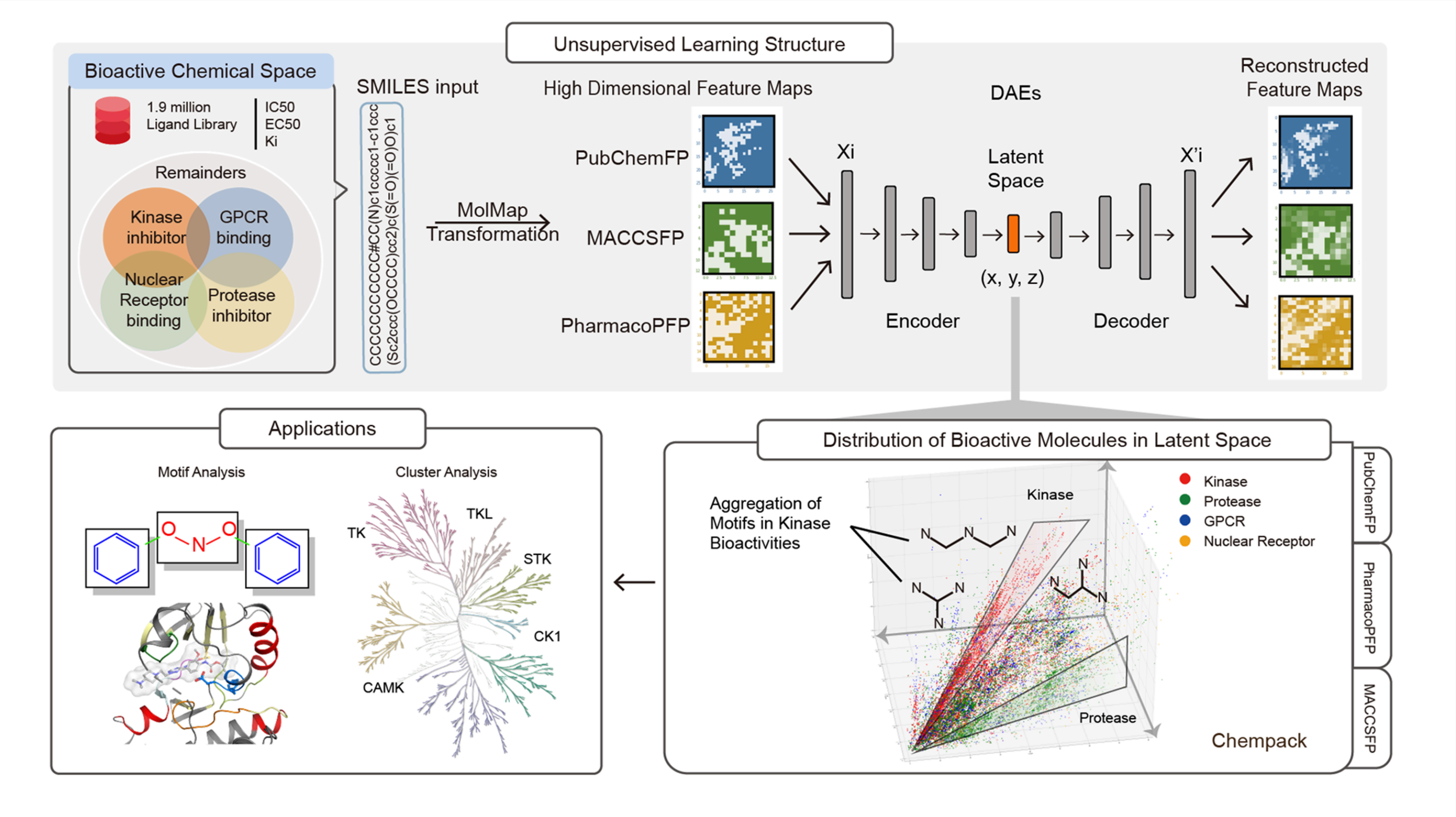

Figure 1

Workflow of MolF-DAEs for unsupervised clustering in construction, visualization, and analysis of biomolecules. Symmetric DNNs are used to build MolF-DAEs. 1.9 million biomolecule data is used as an example, but the strategy is applicable to 3D non-destructive visualisation of other high dimensional data. Five categories of molecules are labeled according to target types (kinase-inhibitor, GPCR-binding, nuclear receptor-binding, protease-inhibitor, and remainders). SMILES are transformed into three types of Fingerprint FMaps. The molecular coordinates in 3DLSpace are visualized using the 3D visualization software, Chempack. The 3D clustering is utilized for downstream applications including motif analysis and bioactivity-related cluster analysis.

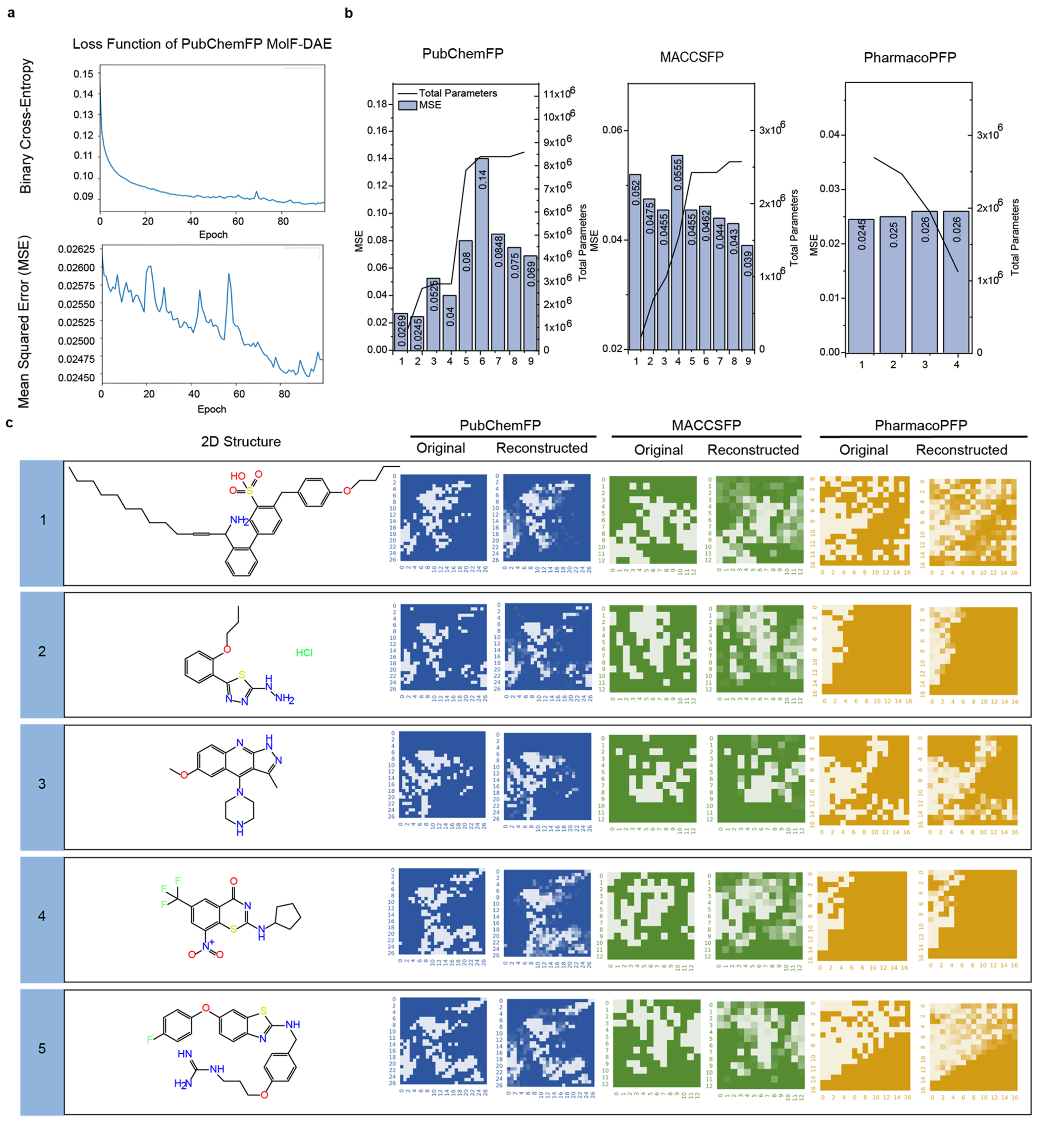

Figure 2

Fully-connected deep autoencoder shows high accuracy in reconstructing Fmaps. (a) Training process of the MSE models. The trend of the loss function in first stage (binary cross entropy) (left) and second stage (MSE) (right) process. (b) Parameter usage and corresponding optimal MSE within different molecular fingerprint channels. (c) Reconstruction effect of Fmaps. Three pairs of Fmaps are shown in different colors. The left represents the original molecular fingerprint images, while the right represents the reconstructed 2D fingerprint Fmaps. White dots denote values of 1, while colored dots (blue: PubChemFP, green: MACCSFP, yellow: PharmacoPFP) represent values of 0, with darker colors indicating closer proximity to 1.

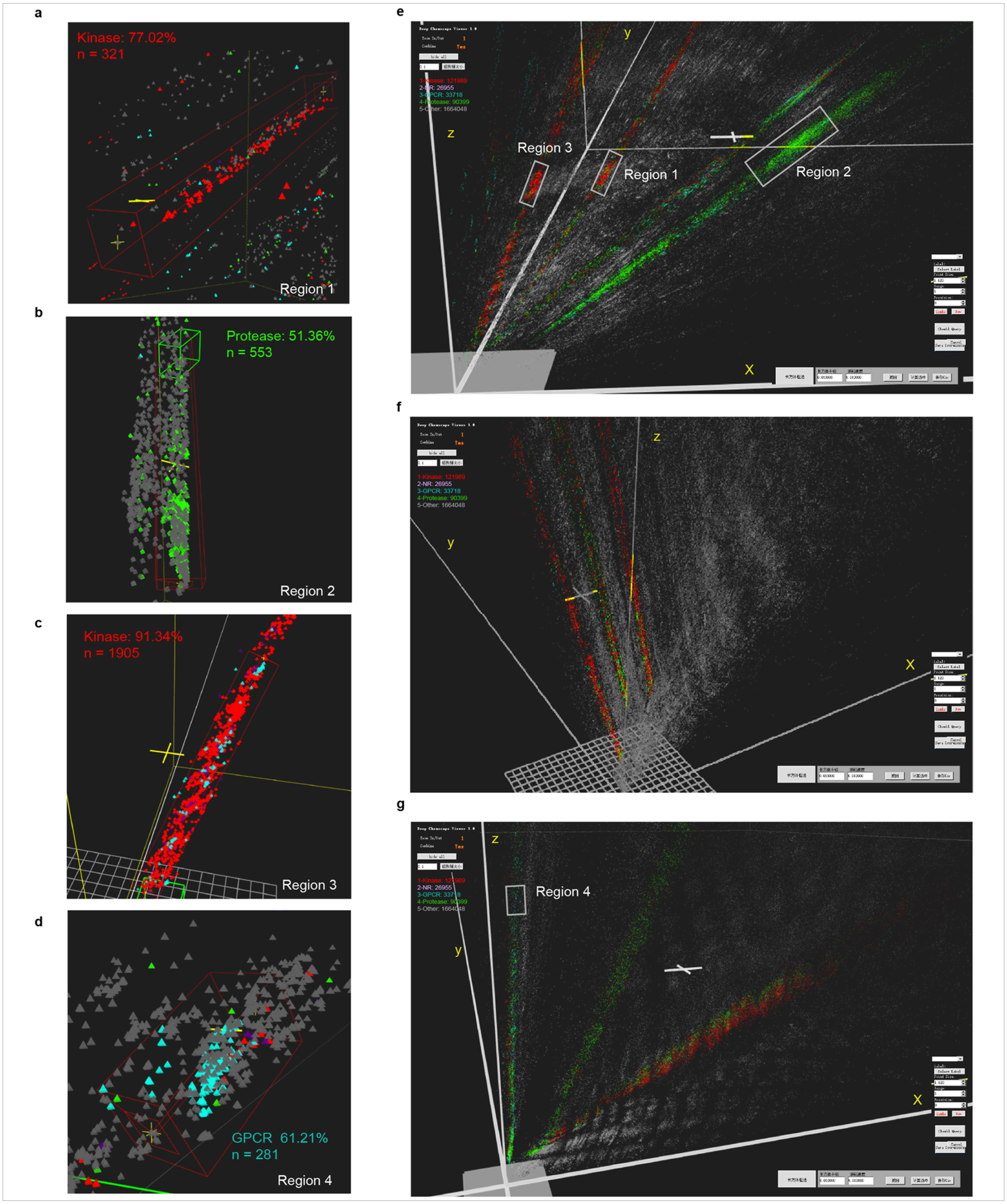

Figure 3

The visualization of biomolecules within 3DLSpace, facilitated by the intuitive interface of the Chempack software. (a-c) Distribution of 1.9 million molecules in (a) PubChemFPM (b) MACCSFPM and (c)

PharmacoPFPM. (d-g) Distribution of biomolecules in 3DLSpace and target labeling cluster to four target types (Red: kinase, green: protease, Cyan: GPCR). Molecules in representative clusters are colored to the target component.

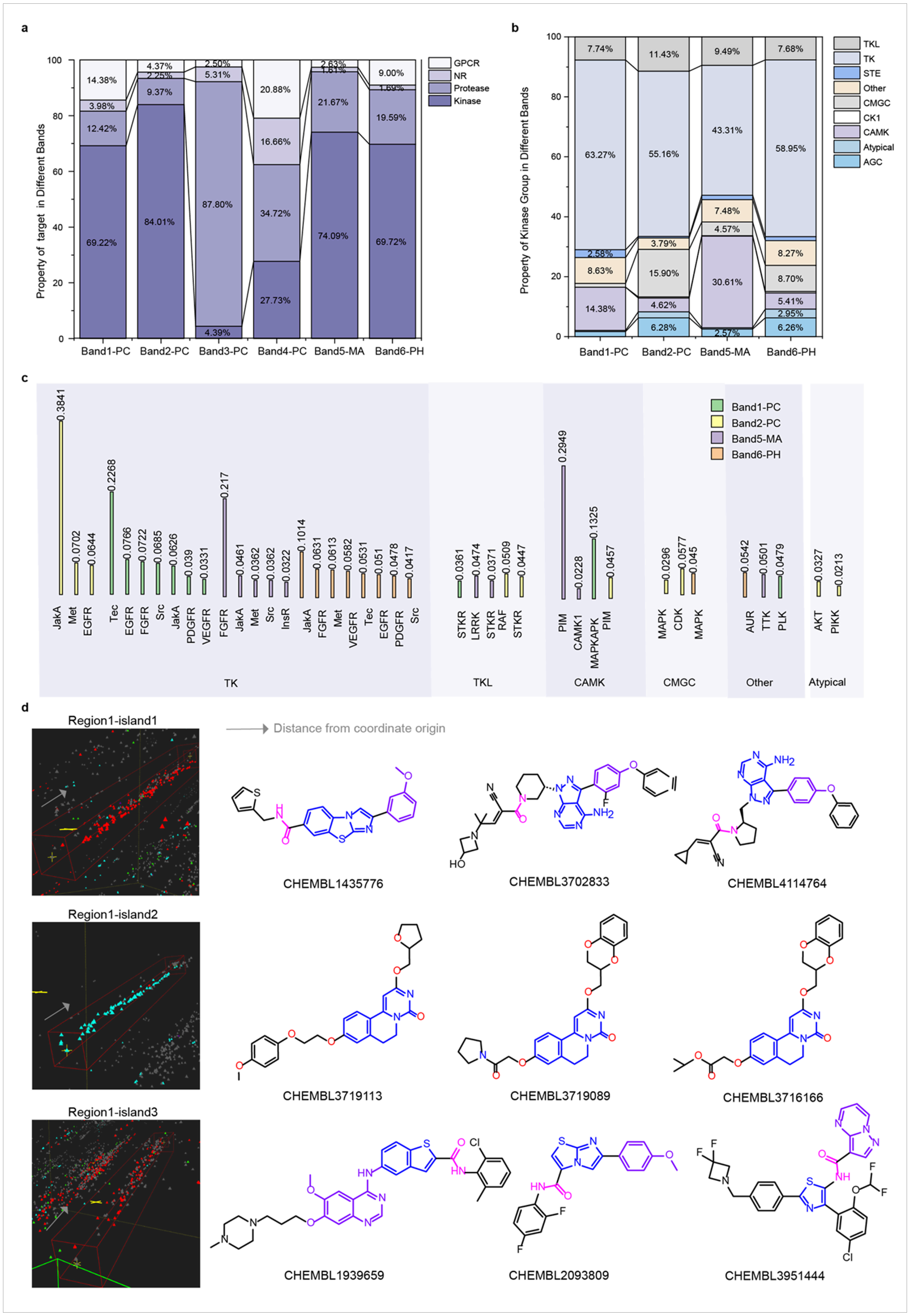

Figure 4

The overall molecule feature distributions and differences in three models in MolF-DAEs. (a) The properties of biomolecules in six representative bands without the gray points. (b) The proportions of all

20 kinase groups within the four kinase-enriched bands 1, 2, 5, and 6. (c) The proportions of the top 10 kinase families within the four kinase-enriched bands 1, 2, 5, and 6. (d) Kinase and GPCR islands in kinase-enriched band 1 and randomly chosen molecule structures and their core elements.

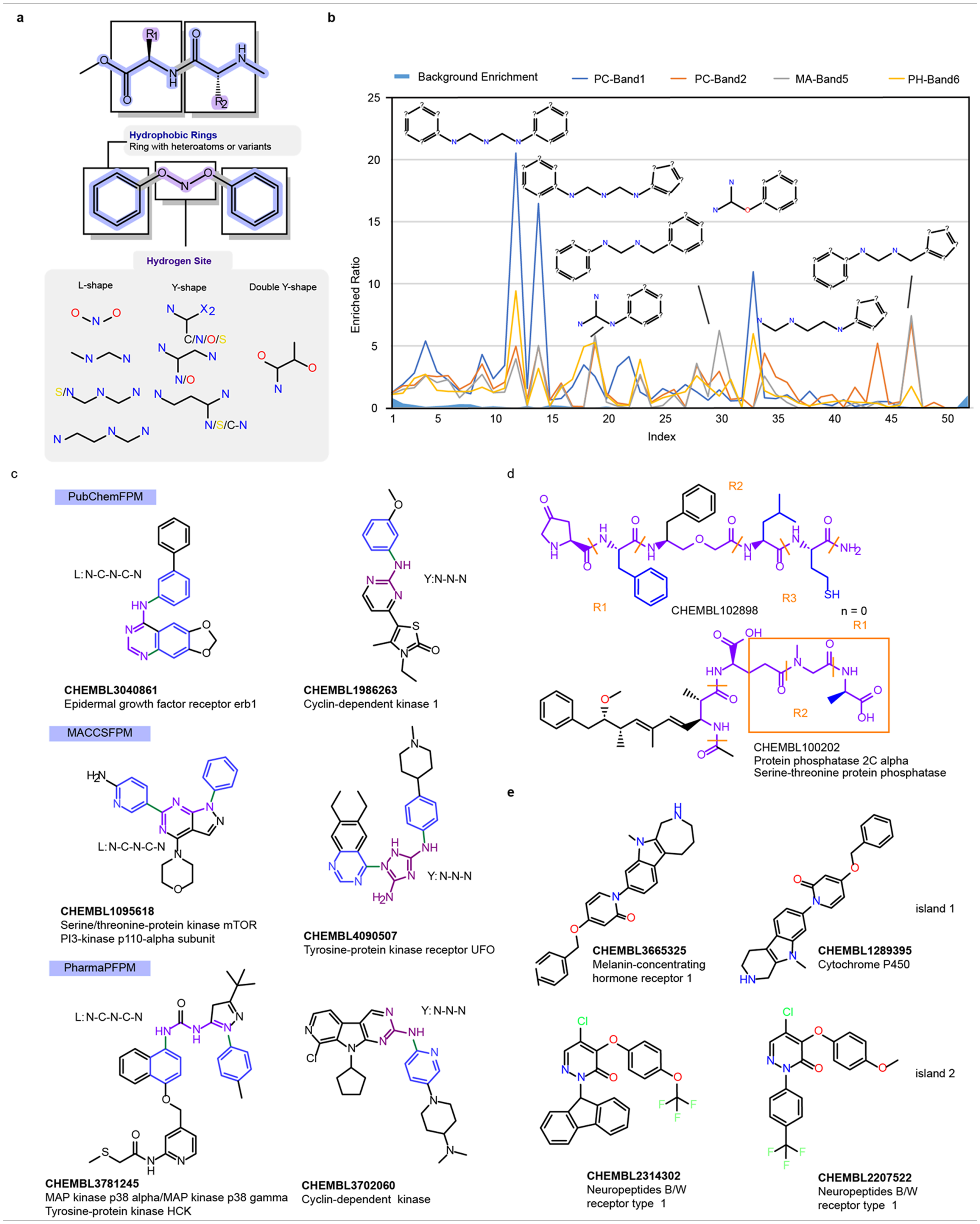

Figure 5

The regularity of substructures explains the patterns of band clustering. (a) Two Models and their substructure composition elements. (b) Substructure enrichment in the 4 kinase-enriched bands compared to the background. (c) Samples and substructure in kinase enriched bands in three models. (Top: PubChemFPM, middle: MACCSFPM, down: PharmacoPFPM) (d) Molecules colored with amide bond and its variants in protease bands. (e) GPCR enriched islands colored by atomic type.

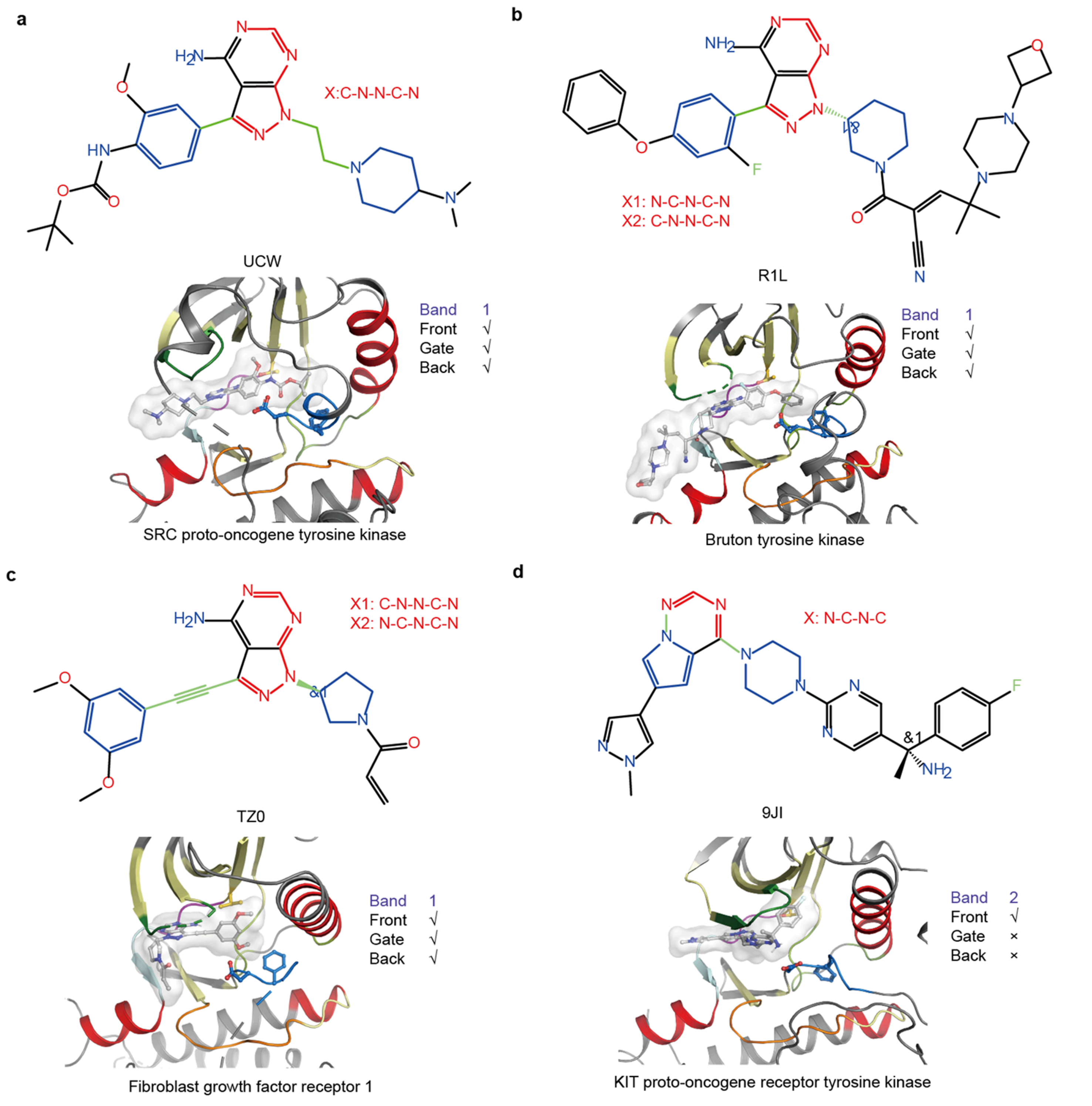

Figure 6

Binding sites of molecules in kinase-enriched bands in PubChemFPM. (a-c) Kinase inhibitors and binding pockets in island 1 in kinase-enriched band 1. Combination pattern and pocket annotation is obtained from KLIFS[50]. (d) Kinase inhibitor and binding pocket in kinase-enriched band 2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supptable16.docx
- supptable716.xlsx
- suppfigure121.docx
- supplementFig22.pdf
- pharmacophore.mp4