Supporting Information

Spectra-Descriptor-Based Machine Learning for Predicting Protein-Ligand Interaction

Cheng Chen,[#,1] Ledu Wang,[#,1] Yi Feng,[#,1] Wencheng Yao,[2] Jiahe Liu,[1] Zifan Jiang,[1] Luyuan Zhao,[1] Letian Zhang,[1] Jun Jiang[1] and Shuo Feng[*,1]

[1] *State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China.*
[2]*MOE Key Laboratory of Resources and Environmental System Optimization, College of Environmental Science and Engineering, North China Electric Power University, Beijing 102206, China*

# C.C. L.W. and Y.F. contributed equally to this work.

*To whom correspondence should be addressed.
Email: sfeng18@ustc.edu.cn (S.Feng)

Single-Target Performance of VS2Net

After obtaining the VS2Net model, we tested its classification ability across different targets. As depicted in the **Figure S1**, the constructed model exhibited remarkable predictive power for the 82 distinct protein targets and their corresponding datasets of active molecules and decoys from the DUD-E database. For the vast majority of targets, the model demonstrated exceptional performance. This result underscores the model's precision and robustness in predicting protein-ligand interactions.

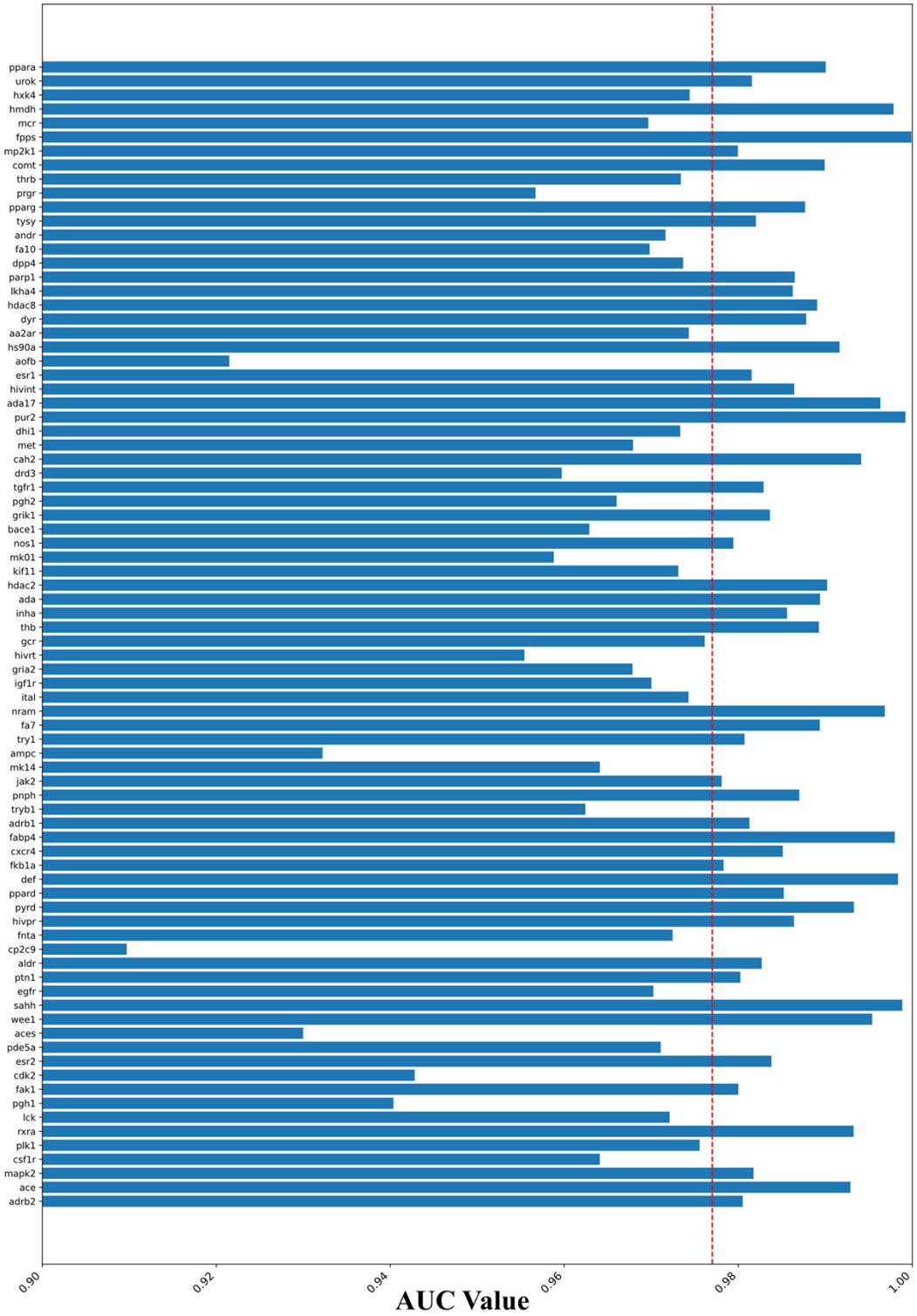

Figure S1. AUC values of the VS2Net model for 82 protein targets in DUD-E dataset.

Target Clustering and Ligands Similarity Evaluation

To explore the potential structural associations among targets represented by the

FISD sequence and validate its effectiveness in ligand recognition, we first conducted a K-means clustering analysis on the 82 known targets from the DUD-E dataset, along with two additional protein targets (SARS-CoV-2 Spike protein and Alzheimer's disease-related Tau protein), with a preset number of clusters set to eight. The targets were clustered using their FISD features as input. To visually represent the clustering results, we employed Principal Component Analysis (PCA) to reduce the dimensionality of the high-dimensional FISD feature space. The result is shown in **Figure S2**.

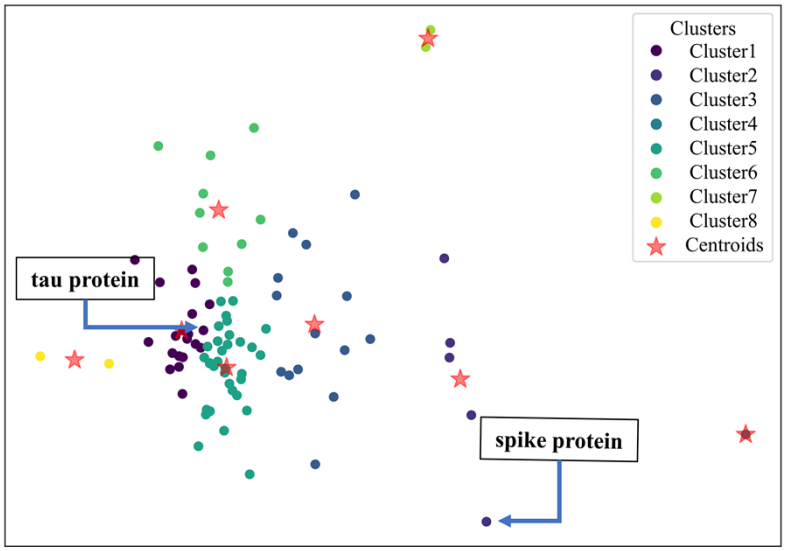

Figure S2. The clustering results for 82 DUD-E targets and 2 additional targets, red star demonstrates the center position of each cluster, Tau protein and Spike protein are pointed out .

Furthermore, we evaluated the similarity of ligand molecules within each cluster and across different clusters. Specifically, molecular similarity was calculated based on their Morgan fingerprints, and the Dice similarity coefficient was used to measure the similarity between two molecules. The ligand similarity between targets was determined by averaging the Dice similarity scores across all possible ligand. As depicted in the **Figure S3**, we observed that the intra-cluster ligand similarity was generally higher than the inter-cluster similarity. This finding preliminarily validates our hypothesis that targets with similar FISD representations may share similar ligand-binding properties, thereby supporting the effectiveness of clustering analysis in reflecting protein-ligand interactions.

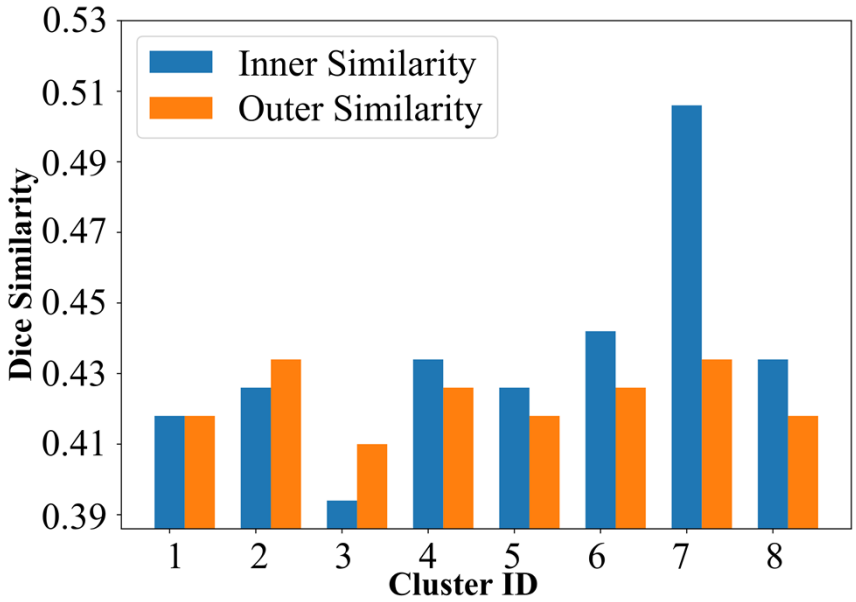

Figure S3. The display plot of inner and outer similarity of each cluster's corresponding ligand library.

Compound Atom Features	
atom type	C, N, O, S, F, P, Cl, Br, Unknown
number of heavy neighboring atoms	0, 1, 2, 3, 4, more
formal charge	-3, -2, -1, 0, 1, 2, 3, extreme
hybridization type	s, sp, sp^2, sp^3, sp^3d, sp^3d^2, other
ring	calculated by RDkit
aromatic	calculated by RDkit
atomic mass	calculated by RDkit
vdw radius	calculated by RDkit
covalent radius	calculated by RDkit

Table S1. detailed description of molecular graph features

Examples of MLMS in predicting DUD-E molecules

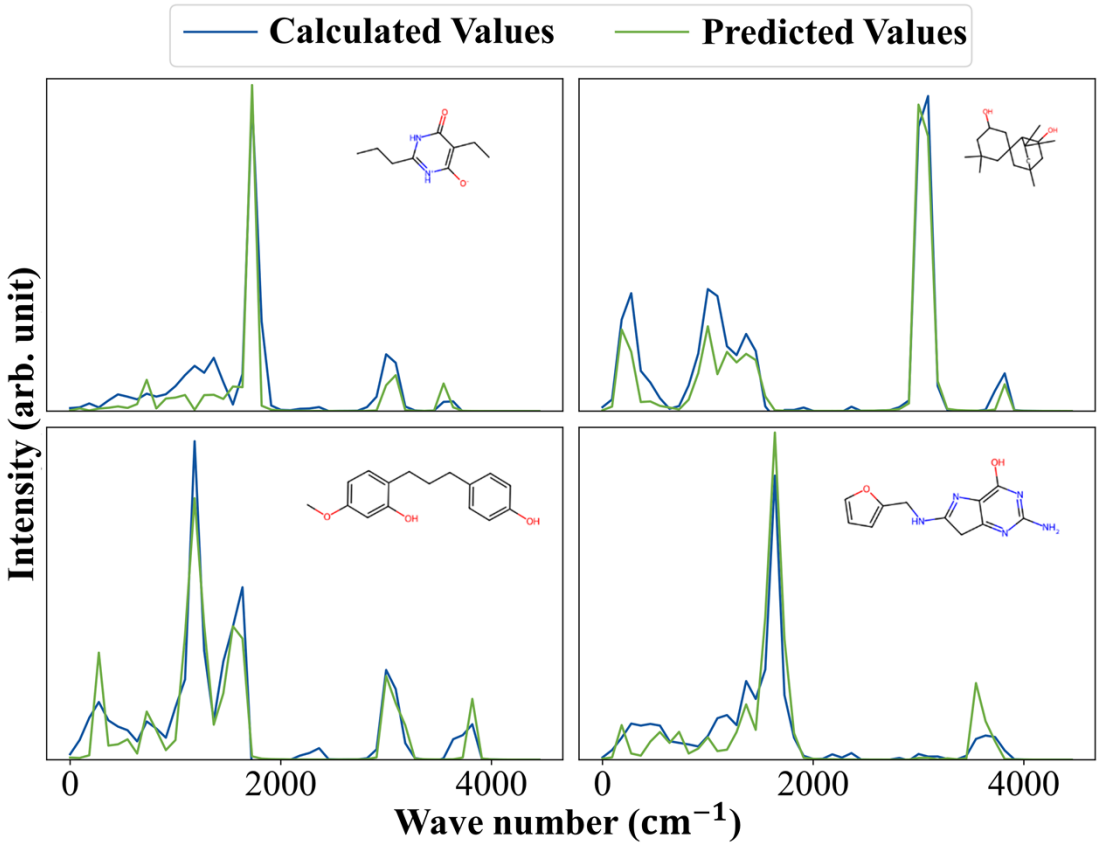

Figure S4. Several examples of MLMS predicted FISD (green line) and their corresponding DFT calculated FISD (blue line).

Related DUD-E Subsets

Prior to training VS2Net with over one million molecules from the DUD-E dataset, we conducted a detailed subdivision of the dataset and preliminary experiments using a model architecture similar to VS2Net. The specific steps we followed are outlined below:

Firstly, we selected 17 protein targets from DUD-E, each with at least 20 active molecules, where the active molecules possessed less than 20 heavy atoms. During preprocessing, we selected 17 eligible target proteins based on specific criteria (e.g., ≥20 active ligands, ≤20 non-hydrogen atoms, and containing only C, O, N, F elements), and the selected target proteins can be seen in **Table S2**.

Target	PDB	Description
ACES	1e66	Acetylcholinesterase
ADA	2e1w	Adenosine deaminase
ANDR	2am9	Androgen Receptor
AOFB	1s3b	Monoamine oxidase B
CDK2	1h00	Cyclin-dependent kinase 2
DPP4	2i78	Dipeptidyl peptidase IV

DRD3	3pbl	Dopamine D3 receptor
ESR1	1sj0	Estrogen receptor alpha
SAHH	1li4	Adenosylhomocysteinase
ESR2	2fsz	Estrogen receptor beta
GRIA2	3kgc	Glutamate receptor ionotropic, AMPA 2
NOS1	1qw6	Nitric-oxide synthase, brain
PNPH	3bgs	Purine nucleoside phosphorylase
GRIK1	1vso	Glutamate receptor ionotropic kainate 1
PARP1	3l3m	Poly [ADP-ribose] polymerase-1
PGH1	2oyu	Cyclooxygenase-1
PGH2	3ln1	Cyclooxygenase-2

Table S2. 17 selected target proteins

We extracted elemental compositions and atom counts from SMILES strings using RDKIT, converted selected active and 100 random decoy molecules (matching active ligand criteria) to suitable format for DFT calculations with OpenBabel. For each target, the randomly sampled 100 inactive molecules in line with the predefined criteria are pairing with the active molecules to form the overall dataset. This resulted in a total of 2200 molecules. Subsequently, we employed Density Functional Theory (DFT) for structural optimization and computed vibrational frequencies to obtain DFT-FISD values. Based on these data, we partitioned the set into training, validation, and test sets (1760/330/110), successfully training a model with good classification performance, as shown in **Figure S5**.

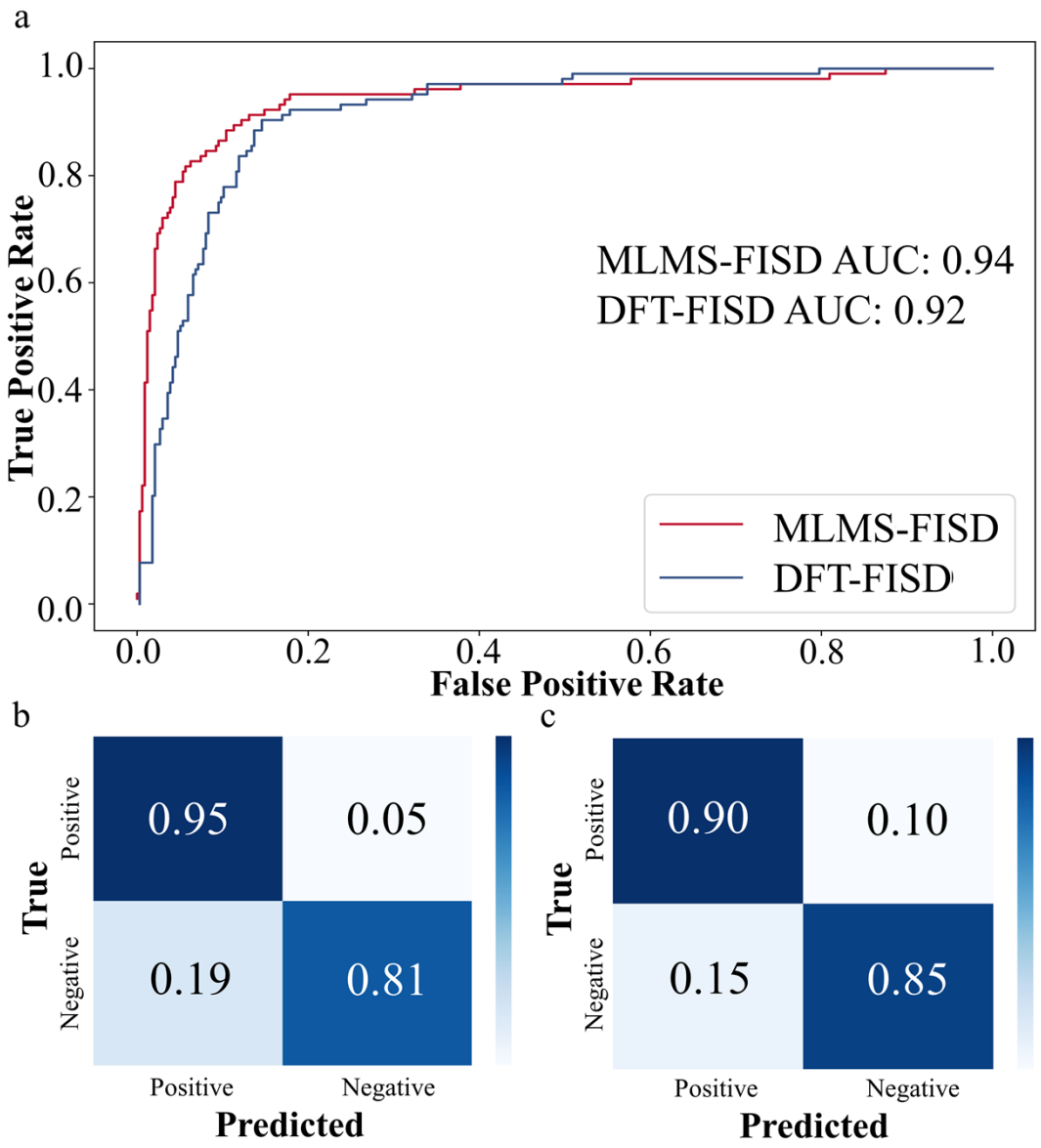

Figure S5. (a) ROC curve of the VS2Net for the DUD-E 2200-set with FISD generated by DFT (blue line) and generated by MLMS (red line). (b) Visualization of the confusion matrix of the VS2Net for the DUD-E 2200-set with FISD generated by MLMS. (c) Visualization of the confusion matrix of the VS2Net for the DUD-E 2200-set with FISD generated by DFT.

Secondly, we recalculated the FISD values for these 2200 molecules using the MLMS model and reassigned the subsets accordingly. The results indicated that the classification performance of the model trained on MLMS-derived FISD values was comparable to that based on DFT-FISD, validating the effectiveness of MLMS-derived FISD in virtual screening tasks, which can be seen in **Figure S5**.

To further examine the applicability of the MLMS method under broader conditions, we expanded the dataset to include all active and inactive molecules (302365 molecules) associated with the aforementioned 17 targets. For each target, we randomly selected 80% of the active molecules corresponding to the target and 10

times as many decoy molecules as the active molecules as the training set (42504 molecules), 10% of the active molecules and 10 times as many decoy molecules as the active molecules as the validation set (5247 molecules), and all the remaining molecules as the test set (254614 molecules). We computed their FISD values using MLMS and trained a new classification model, which demonstrated excellent performance, as shown in **Figure S6**. This supported our hypothesis that MLMS can provide effective FISD values even for molecules with an increased number of heavy atoms and the provided FISD can be used to fulfill the protein-ligand interaction identification task.

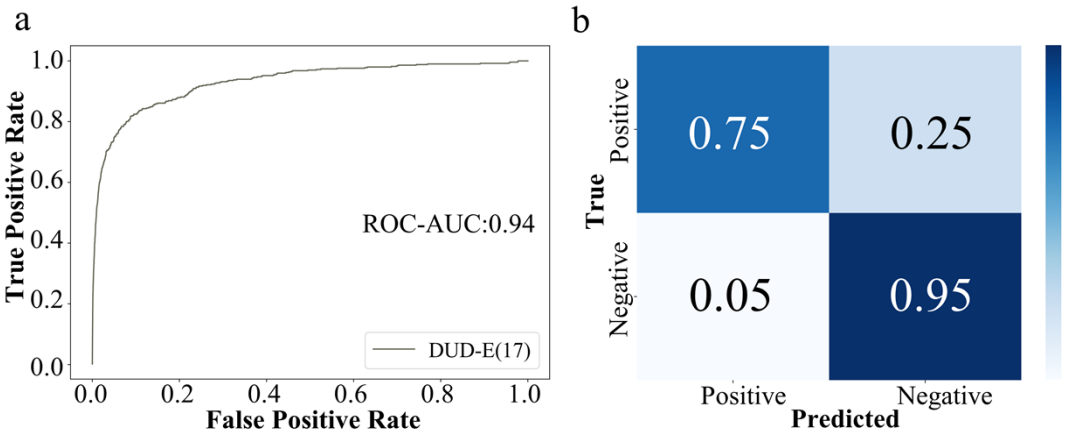

Figure S6. (a) ROC curve of the VS2Net for the DUD-E 17 test set with FISD generated by MLMS (roughly 240k molecules). (b) Visualization of the confusion matrix of the VS2Net for the DUD-E 17 test set.

Next, we conducted transfer learning experiments based on the aforementioned classification model, and the result can be seen in **Figure S7**. We selected one additional target from the DUD-E dataset (HIVPR) and trained the model using its corresponding molecular data. During training, we gradually increased the data volume while using the remaining data for validation and testing to observe changes in model performance. The results showed that a significant improvement in classification ability could be achieved with only a small amount of data, and as the data volume continued to increase, model performance stabilized. To achieve optimal performance, approximately 60 active and 600 inactive data points were required for training.

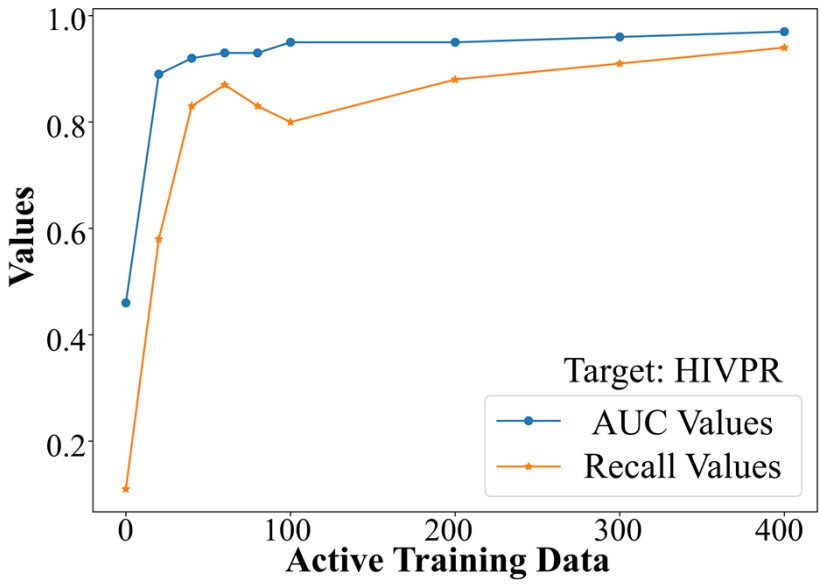

Figure S7. Visualization of the transfer learning results. The training data contains corresponding number of active training data and decoy training data which is ten times bigger than the active training data. The orange line demonstrates the value fluctuation of the model's recall values, and the blue line shows the AUC values.

Finally, we trained the VS2Net model using all data from the 82 targets in the DUD-E dataset，the dataset was partitioned as the way consistent with the previous methodology..

Enrichment Factor (EF/RE)

In evaluating the classification results of the VS2Net model, we adhered to the established research conventions by adopting the Enrichment Factor (EF) as one of the key metrics. While EF is occasionally abbreviated as RE in some literature, both abbreviations refer to the same concept, which quantifies the model's ability to enrich active molecules in the early stages of the screening process. To maintain consistency, this paper uniformly uses EF as the abbreviation.

Specifically, the EF score is defined as the ratio of the True Positive Rate (TPR) to the False Positive Rate (FPR) at a particular FPR threshold. This metric directly reflects the model's capability to correctly identify target molecules while controlling the false recognition rate. In this study, we selected FPR thresholds of 0.5%, 1%, 2%, and 5%, based on normally-used criteria of previous work, aiming to comprehensively and meticulously assess the enrichment performance of the VS2Net model under varying degrees of stringency.

By evaluating the model's performance at these specific FPR thresholds, we can gain insights into how effectively the VS2Net model prioritizes active molecules over decoys, even at very low error rates. This information is crucial for guiding the selection of optimal screening strategies and thresholds in practical drug discovery applications.

Model Parameters

The MLMS model comprises three primary components, two of which are core components constructed upon Graph Convolutional Networks (GCNs), exhibiting a high degree of similarity with the primary difference lying in the employed loss functions: one utilizes Mean Squared Error (MSE) loss, while the other employs Cosine Embedding (CosEmbedding) loss. Both components accept molecular graphs as input, proceed through five layers of GCNConv for feature extraction, and then employ global mean pooling and global max pooling strategies to aggregate graph-level features into fixed-dimensional representations. These aggregated features are concatenated and passed through two linear layers to the output layer, ultimately mapping the molecules into a 50-dimensional output space. During this process, ReLU activation functions are applied between the linear layers, and the Adam optimizer is configured with a learning rate of 0.0001 and a weight decay rate of 0.0005 for model optimization. The outputs of the two GCN components are then concatenated into a 100-dimensional vector, serving as input for subsequent processing stages. This vector undergoes further processing through a deep network consisting of four hidden layers, each set to 512 dimensions. Ultimately, the network outputs a 50-dimensional result, with this stage also employing MSE loss, ReLU activation functions, and the Adam optimizer (with a learning rate of 0.0001 and a weight decay of 0.0005) for training.

The VS2Net model receives a 100-dimensional input vector, which is concatenated from 50-dimensional FISD feature vectors of both the molecule and the protein. The model architecture encompasses five hidden layers, sequentially mapping the input from 100 dimensions to 128, 256, 128, 64, and finally 32 dimensions. At the output layer, the model generates a one-dimensional prediction. During training, Binary Cross-Entropy (BCE) loss function is employed, in conjunction with ReLU and Sigmoid activation functions to enhance the model's expressive power. The Adam optimizer is chosen, configured with a learning rate of 0.0001 and a weight decay rate of 0.0005.

All models were trained on an NVIDIA GeForce RTX 3090 GPU. For the MLMS model, the three training processes were carried out over 2000, 2000 and 100 epochs, whereas the VS2Net model underwent 500 epochs of training. During the training phase, the performance on the validation set was monitored, and the best-performing model was retained as the final model.

Descriptor	Descriptor
'MaxAbsEStateIndex'	'NHOHCount'
'MaxEStateIndex'	'NOCount'
'MinAbsEStateIndex'	'NumAliphaticCarbocycles'
'MinEStateIndex'	'NumAliphaticHeterocycles'
'qed'	'NumAliphaticRings'

'SPS'
'MolWt'
'HeavyAtomMolWt'
'ExactMolWt'
'NumValenceElectrons'
'NumRadicalElectrons'
'MaxPartialCharge'
'MinPartialCharge'
'MaxAbsPartialCharge'
'MinAbsPartialCharge'
'BCUT2D_MWHI'
'BCUT2D_MWLOW'
'BCUT2D_CHGHI'
'BCUT2D_CHGLO'
'BCUT2D_LOGPHI'
'BCUT2D_LOGPLOW'
'BCUT2D_MRHI'
'BCUT2D_MRLOW'
'AvgIpc'
'BalabanJ'
'BertzCT'
'Chi0'
'Chi0n'
'Chi0v'
'Chi1'
'Chi1n'
'Chi1v'
'Chi2n'
'Chi2v'
'Chi3n'
'Chi3v'
'Chi4n'
'Chi4v'
'HallKierAlpha'
'Kappa1'
'Kappa2'
'Kappa3'
'LabuteASA'
'PEOE_VSA1'
'PEOE_VSA10'
'PEOE_VSA11'
'PEOE_VSA12'
'PEOE_VSA13'
'PEOE_VSA14'
'NumAromaticCarbocycles'
'NumAromaticHeterocycles'
'NumAromaticRings'
'NumHAcceptors'
'NumHDonors'
'NumHeteroatoms'
'NumRotatableBonds'
'NumSaturatedCarbocycles'
'NumSaturatedHeterocycles'
'NumSaturatedRings'
'RingCount'
'MolLogP'
'MolMR'
'fr_Al_COO'
'fr_Al_OH'
'fr_Al_OH_noTert'
'fr_ArN'
'fr_Ar_COO'
'fr_Ar_N'
'fr_Ar_NH'
'fr_Ar_OH'
'fr_COO'
'fr_COO2'
'fr_C_O'
'fr_C_O_noCOO'
'fr_C_S'
'fr_HOCCN'
'fr_Imine'
'fr_NH0'
'fr_NH1'
'fr_NH2'
'fr_N_O'
'fr_Ndealkylation1'
'fr_Ndealkylation2'
'fr_Nhpyrrole'
'fr_SH'
'fr_aldehyde'
'fr_alkyl_carbamate'
'fr_alkyl_halide'
'fr_allylic_oxid'
'fr_amide'
'fr_amidine'
'fr_aniline'
'fr_aryl_methyl'

'PEOE_VSA2'
'PEOE_VSA3'
'PEOE_VSA4'
'PEOE_VSA5'
'PEOE_VSA6'
'PEOE_VSA7'
'PEOE_VSA8'
'PEOE_VSA9'
'SMR_VSA1'
'SMR_VSA10'
'SMR_VSA2'
'SMR_VSA3'
'SMR_VSA4'
'SMR_VSA5'
'SMR_VSA6'
'SMR_VSA7'
'SMR_VSA8'
'SMR_VSA9'
'SlogP_VSA1'
'SlogP_VSA10'
'SlogP_VSA11'
'SlogP_VSA12'
'SlogP_VSA2'
'SlogP_VSA3'
'SlogP_VSA4'
'SlogP_VSA5'
'SlogP_VSA6'
'SlogP_VSA7'
'SlogP_VSA8'
'SlogP_VSA9'
'TPSA'
'EState_VSA1'
'EState_VSA10'
'EState_VSA11'
'EState_VSA2'
'EState_VSA3'
'EState_VSA4'
'EState_VSA5'
'EState_VSA6'
'EState_VSA7'
'EState_VSA8'
'EState_VSA9'
'VSA_EState1'
'VSA_EState10'
'fr_azide'
'fr_azo'
'fr_barbitur'
'fr_benzene'
'fr_benzodiazepine'
'fr_bicyclic'
'fr_diazo'
'fr_dihydropyridine'
'fr_epoxide'
'fr_ester'
'fr_ether'
'fr_furan'
'fr_guanido'
'fr_halogen'
'fr_hdrzine'
'fr_hdrzone'
'fr_imidazole'
'fr_imide'
'fr_isocyan'
'fr_isothiocyan'
'fr_ketone'
'fr_ketone_Topliss'
'fr_lactam'
'fr_lactone'
'fr_methoxy'
'fr_morpholine'
'fr_nitrile'
'fr_nitro'
'fr_nitro_arom'
'fr_nitro_arom_nonortho'
'fr_nitroso'
'fr_oxazole'
'fr_oxime'
'fr_para_hydroxylation'
'fr_phenol'
'fr_phenol_noOrthoHbond'
'fr_phos_acid'
'fr_phos_ester'
'fr_piperdine'
'fr_piperzine'
'fr_priamide'
'fr_prisulfonamd'
'fr_pyridine'
'fr_quatN'

'VSA_EState2'	'fr_sulfide'
'VSA_EState3'	'fr_sulfonamd'
'VSA_EState4'	'fr_sulfone'
'VSA_EState5'	'fr_term_acetylene'
'VSA_EState6'	'fr_tetrazole'
'VSA_EState7'	'fr_thiazole'
'VSA_EState8'	'fr_thiocyan'
'VSA_EState9'	'fr_thiophene'
'FractionCSP3'	'fr_unbrch_alkane'
'HeavyAtomCount'	'fr_urea'

Table S3. The list of 206 descriptors that make up Cheminfo-D, labeled by their names in RDkit.

Molecular Dynamics

Simulation Details

First, we screened the molecules and conformations with the highest docking scores and generated the required topology and force field files using the SobTop tool.[1] The parameters for these files were derived from the standard Amber GAFF force field 4.[2] For the protein portion, we used the pdb2gmx tool built into the GROMACS 2022 version to generate the topology file, with parameters taken from the Amber99sb-ildn force field.[3, 4]

Based on the conformations obtained from the docking results, we constructed the complex structure files and defined the simulation box. Subsequently, we filled the box with water molecules (using the three-point water model spc216) and added ions to maintain the charge balance of the system. To prepare for the simulation, we performed 5,000 steps of energy minimization (EM) on the complex and conducted a 500 ps molecular dynamics (MD) run using the NVT ensemble at T=300 K to achieve equilibration. Then, using the NPT ensemble, we performed MD simulations at T=300 K and P=1 bar for all model systems during the production stage. For the Spike protein and its ligand molecules, we generated a 100 ns trajectory, and for the Tau protein and its ligand molecules, we generated a 50 ns trajectory for subsequent MD analysis.

We simulated six systems, which are as follows: tau-5 with Tau protein, tau-active with Tau protein, tau-inactive with Tau protein, spike-4 with Spike protein, spike-active with Spike protein, and spike-inactive with Spike protein. Among them, tau-5 and spike-4 are predicted results from VS2Net and are sourced from the DUD-E database. tau-active is obtained from BindingDB[5], and spike-active is sourced from work by Timoteo et. al.[6] tau-inactive and spike-inactive are randomly selected molecules. The specific Spike protein used is the Spike protein (PDB ID: 7lm9), and the Tau protein used is the Tau protein (PDB ID: 8q96).

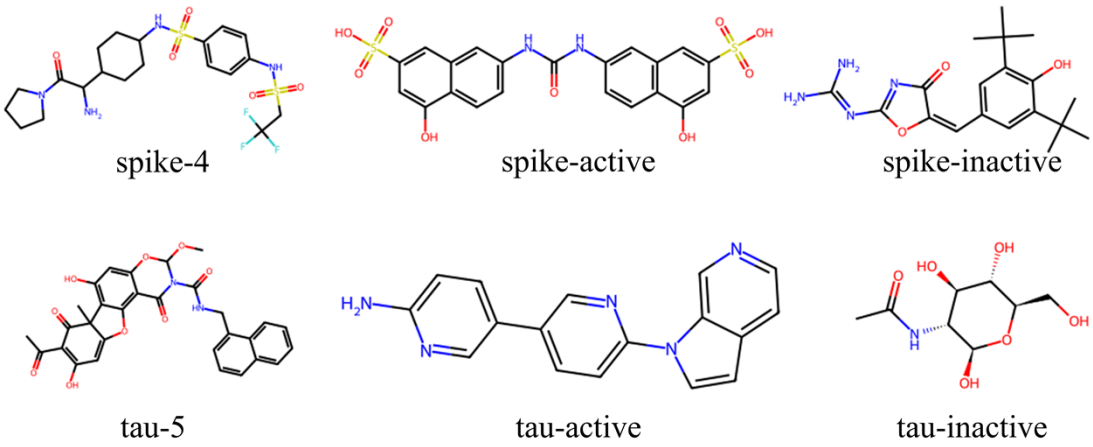

Figure S8. Molecules used in MD simulations. spike-4 and spike-inactive are from DUD-E dataset, and spike-active is from a published journal article. tau-5 and tau-inactive are from DUD-E dataset, and tau-active is from BindingDB.

Simulation Results

To further validate the binding ability of the ligands obtained through our method with the target proteins, after performing molecular docking using AutoDock Vina, we selected the molecules with the highest affinity and their corresponding basic conformations for more detailed kinetic simulations. For the protein-ligand complexes corresponding to these two cases, we separately conducted kinetic simulations and achieved conformational equilibration during MD simulations under the NPT ensemble. We collected the RMSD of the ligands during the simulations, as shown in **Figure S9** and **Figure S10**:

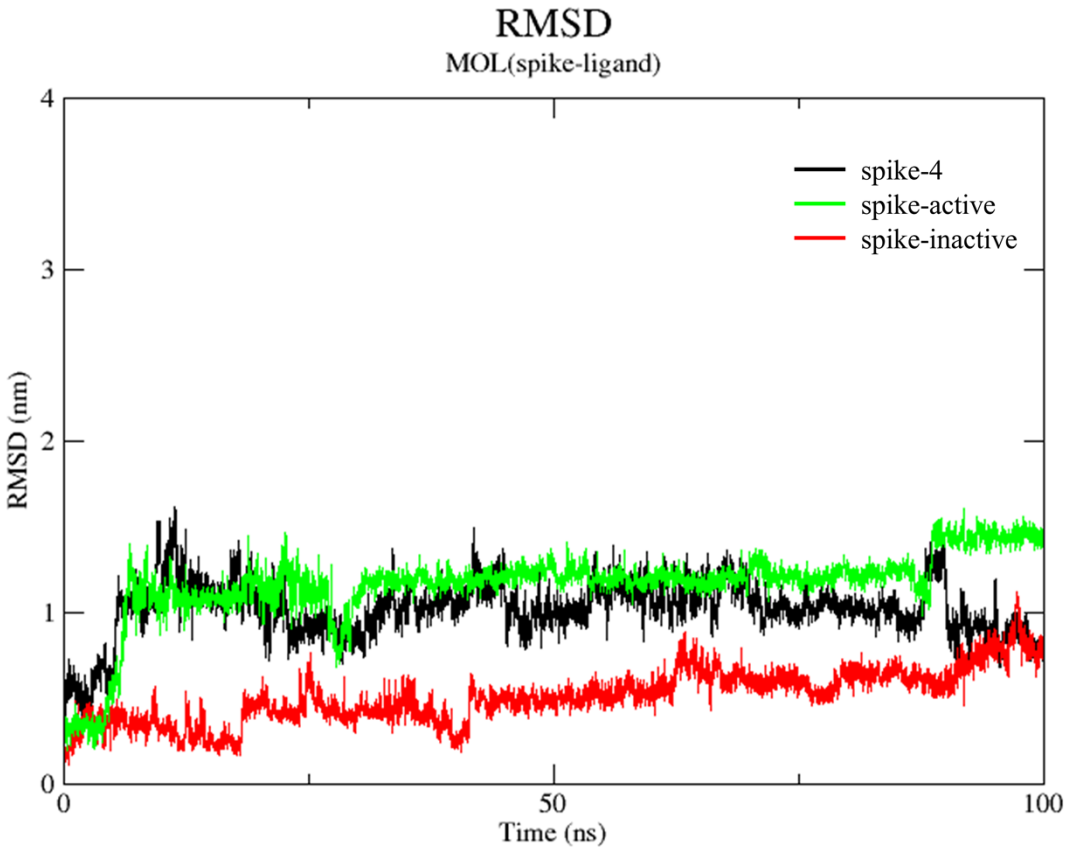

Figure S9. MD results for Spike protein and its corresponding molecules' RMSD

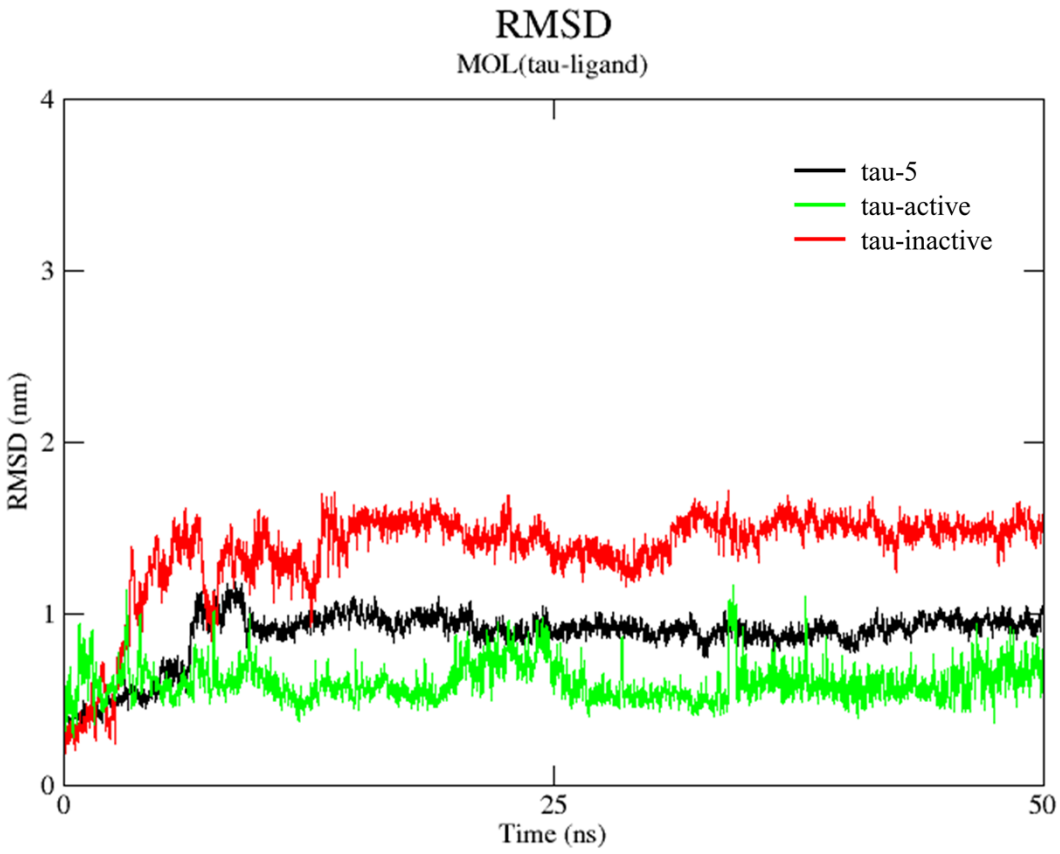

Figure S10. MD results for Tau protein and its corresponding molecules' RMSD

The RMSD value represents the deviation of the structure at a certain time point from the initial conformation, reflecting the stability of the system during the simulation. The above figure shows the RMSD changes of the ligand molecules in each system during the simulation. Since the initial conformations for the simulations were obtained through AutoDock, significant conformational changes occurred in the molecules during the initial stages of MD. However, the stabilization of ligand RMSD in the later stages of the simulation indicates that the system has become stable, with minimal conformational changes and the system's energy reaching a lower, stable state.

To further assess the binding situation between the molecules and proteins, we also calculated the binding free energy. In addition to conducting MD simulations and related analyses on the results obtained from VS2Net, we also selected one experimentally validated ligand and one randomly picked non-ligand for each of the two target proteins, performed molecular docking and MD simulations, and compared the results with those from VS2Net. We selected a stable trajectory from the later stages of the simulation and use the MM/PBSA method to do the calculation and analysis of binding free energy. The results showed that both the complexes obtained through VS2Net, tau-5 with Tau protein and spike-4 with Spike protein, exhibited very strong binding free energy compared to the experimentally validated ligands tau-

active and spike-active. This indicates that they bind stably to the target proteins with high affinity and are indeed potential high-quality candidate ligands. In contrast, the randomly selected tau-inactive and spike-inactive had lower binding energies, suggesting unstable binding and their unsuitability as ligands, which aligns with the results from VS2Net.

Target	Compound	MM/PBSA (kcal/mol)
Tau	tau-5	-56.25
Tau	tau-active	-9.47
Tau	tau-inactive	-0.77
Spike	spike-4	-20.85
Spike	spike-active	-35.08
Spike	spike-inactive	-0.36

Table S4. The MM/PBSA results for each MD system

References

1. T. Lu, Sobtop, Version_1.0, http://sobereva.com/soft/Sobtop (accessed on 16/1/2025), 2025.
2. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *Journal of Computational Chemistry*, 2004, **25**, 1157-1174.
3. M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, 2015, **1-2**, 19-25.
4. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, Improved side-chain torsion potentials for the Amber ff99SB protein force field, *Proteins: Structure, Function, and Bioinformatics*, 2010, **78**, 1950-1958.
5. T. Q. Liu, L. Hwang, S. K. Burley, C. Nitsche, C. Southan, W. P. Walters and M. K. Gilson, BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data, *Nucleic Acids Research*, 2024, DOI: 10.1093/nar/gkae1075.
6. T. Delgado-Maldonado, A. González-González, A. Moreno-Rodríguez, V. Bocanegra-García, A. V. Martinez-Vazquez, E. d. J. de Luna-Santillana, G. Pujadas, G. Rojas-Verde, E. E. Lara-Ramírez and G. Rivera, Ligand- and Structure-Based Virtual Screening Identifies New Inhibitors of the Interaction of the SARS-CoV-2 Spike Protein with the ACE2 Host Receptor, *Pharmaceutics*, 2024, **16**, 613.