Supplementary Information

S.1 Model Details

Here we give a brief description of the geometric deep learning model of ScanNet, which is used in S2Site.

From a given protein structure, the input information is extracted as detailed in Table S3. ScanNet first computes a local reference frame f for every atom using the Point Cloud $P \in \mathbb{R}^{N\times 3}$ and Triplets of Neighborhood Indices $T \in \mathbb{R}^{M\times 3}$. Each index in $T_m = (t_{m1}, t_{m2}, t_{m3})$ refers to an atom 3D coordinates in P. Each $f \in \mathbb{R}^{4\times 3}$ is derived from Equation (S1) to (S4), which represents the local frame's center and along the x, y, and z-axis in the Euclidean space.

$$f_1 = P_{t_{n1}} \tag{S1}$$

$$f_4 = \frac{P_{t_{n3}} - P_{t_{n1}}}{||P_{t_{n3}} - P_{t_{n1}}||} \tag{S2}$$

$$f_3 = \frac{f_4 \times (P_{t_{n2}} - P_{t_{n1}})}{||f_4 \times (P_{t_{n2}} - P_{t_{n1}})||} \tag{S3}$$

$$f_2 = \frac{f_3 \times f_4}{||f_3 \times f_4||} \tag{S4}$$

For each atom, its top K_1 closest neighbors are searched by computing and comparing the Euclidean distances between its frame and other frame centers. The neighborhood feature is formed using the K_1 nearest neighbors' coordinates $C \in R^{K_1\times 3}$ and attributes $D \in \mathbb{R}^{K_1\times 12}$. The neighborhood coordinates convolve with $G = 32$ Gaussian kernels to form a sparse representation of the structure data that is further convolved with trainable filters and attributes for a set of $F = 128$ spatial-chemical feature maps, as shown in equation (S6) and (S5).

$$y_f = \text{ReLU}\left[\mathbf{1}_G^T(W_f^1 \odot (GD))\mathbf{1}_{12} + W_f^2 G\mathbf{1}_G + W_f^b\right],\ f \in [1, F] \tag{S5}$$

$$G_{gk} := G_g(\mu_g, \sigma_g, C_k) = \exp[-\frac{1}{2}(C_k - \mu_g)^T\Sigma_g^{-1}(C_k - \mu_g)],\ g \in [1, G]\ , k \in [1, K_1] \tag{S6}$$

where $W_f^1 \in \mathbb{R}^{G\times 12}$ and $W_f^2 \in \mathbb{R}^{1\times G}$ are trainable weights, $W_f^b \in \mathbb{R}$ is the bias, and $\mathbf{1}_d$ is the vector of ones of length d.

A binary bipartite graph is then built from residues to atoms using the residue and atom-level point clouds. It finds the top K_2 nearest atoms to every residue for the corresponding features, denoted as $y \in \mathbb{R}^{F\times K_2}$. The extracted atom features are down-sampled by a customized attention layer using equation (S7). The layer aims to have a comprehensive feature $N = 64$ extracted from the atoms. The atomic feature is condensed to have the same length L as its sequence.

$$y_n = \left[W_n^3 y \odot \frac{\exp[W_n^4 y - \max_k(W_n^4 y)]}{\exp[W_n^4 y - \max_k(W_n^4 y)]\mathbf{1}_{K_2}}\right]\mathbf{1}_{K_2},\ n \in [1, N] \tag{S7}$$

where $W_n^3 \in \mathbb{R}^{1\times F}$ and $W_n^4 \in \mathbb{R}^{1\times F}$ are trainable linear projection and attention-weighting matrices. $\mathbf{1}_{K_2}$ is the column vector of ones of length K_2.

At the residue level, the residue embedding has been passed through a dropout layer at 0.5 probability, followed by an element-wise dense layer with ReLU activation function to obtain each $res_i \in \mathbb{R}^{1\times 32}$. Each pooled atomic feature $y_i \in \mathbb{R}^{1\times N}$ is residue-wise concatenated to the res_i. The calculation of the frames, neighborhood coordinates, and attributes is then performed at the residue-level frame. The customized attention layer with the softmax activation function serves as a classifier that predicts every residue as binding or non-binding. The cross entropy is used as the loss function to train the model.

S.2 Evaluation criteria

Considering the highly imbalanced nature of the datasets in our work, we assess S2Site's performance using recall, precision, and Matthews correlation coefficient (MCC). The metric calculation formulas are shown below:

$$Recall = TPR = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP (true positive) and TN (true negative) respectively denote the number of residues that are correctly predicted as binding and non-binding; FP (false positive) and FN (false negative) respectively denote the number of residues that are incorrectly predicted as binding and non-binding. Recall measures the proportion of the correctly predicted binding residues over the total binding residues, and precision measures the prediction accuracy from the predicted binding residues. MCC provides a more comprehensive measure of the model performance by taking into account both binding and non-binding residues. However, these metrics can be greatly impacted by the selected threshold. Therefore, threshold-free metrics, the area under the precision-recall curve (AUPRC), and the area under the receiver operating characteristic curve (ROC AUC) are also chosen to evaluate our model performance.

S.3 Training

We apply mini-batch training to conserve computational costs during the training process. The model is trained to minimize the binary cross-entropy loss with the Adam optimizer of learning rate 10^{-3}. A learning rate scheduler and an early stopper are also applied based on the validation cross-entropy. While proteins consist of sequences with varying lengths, GPU-based training requires inputs with fixed dimensions. We present two approaches to form training samples from proteins, catering to different model training that requires different input features. In the case of S2Site training, we adopt the methodology outlined in ScanNet. That is, each sample comprises a greedy concatenation of multiple proteins. To avoid overlapping between concatenated proteins within a sample, we introduce a large translation to the coordinates to make the proteins far from each other. As protein sequences cannot be perfectly partitioned into batches with the specific maximal sequence length L_{max}, we either apply zero-padding to the concatenated sample or truncate the exceeding sequence to achieve a fixed L_{max}. In the baseline methods of U-Net and RCNN, we stack proteins of similar sequence lengths to form a training batch instead of concatenating protein sequences such that the total length of sequences achieves L_{max}. The samples in each batch are padded to the maximum sequence length within the batch. As such, batches could have dynamic batch sizes while minimizing the need for padding and truncating. To balance computational efficiency, generalization of the learned pattern, and the retention of information from larger proteins, L_{max} equals 1024, 2120, and 1485 for the protein-binding site, B-cell epitope, and peptide-binding site predictions, respectively.

For hyper-parameter selection, we use leave-one-out cross-validation in the protein and peptide-binding site predictions while applying 5-fold cross-validation in the B-cell epitope prediction. We also employ transfer learning (with learning rate 10^{-4}) to predict B-cell epitopes and peptide-binding sites to compensate for the limited data and exploit the prior knowledge learned from the large pool of protein-protein interaction data.

S.4 Protein sequence similarity

The splitting of the training and test set for the protein-protein binding task is done by four levels of similarities. Here we provide the sequence similarities between the training set and the four test sets computed by BLOSUM. The average sequence similarities are 24.3%, 19.3%, 19.2% and 18.6% for the four test sets (Test 70%, Test Homology, Test Topology and Test None). The 99% percentage of the similarities are below 37.6%, 37.15%, 24.5% and 24.3% (i.e., 99% percentile) for the four test sets. We will add these statistics to the paper to enhance the reliability of our model's performance assessments. To further ensure the transparency of the transfer learning setting for peptide binding site prediction, we compute the pairwise sequence identities between the training proteins for the protein-binding site model and the test proteins for protein-peptide binding, as shown in Figure S1. Most of the similarities are below 40%, indicating the diversity of the training and test protein sequences.

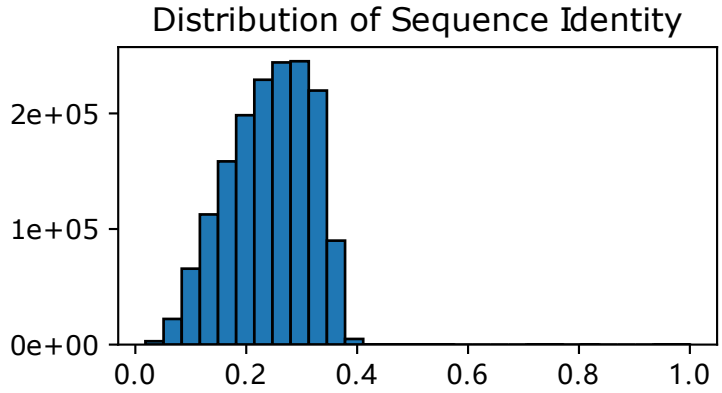

Figure S1: A histogram showing the distribution of the pairwise pairwise sequence identities between the training proteins for the protein-binding site model and the test proteins for protein-peptide binding.

S.5 Supplementary Tables

Table S1: Summary of datasets used in the protein-protein interaction task.

Datasets	Train	Val				Test			
		70%	Homology	Topology	None	70%	Homology	Topology	None
Number of chains	12,773	414	1,458	590	754	554	1,488	915	1,079
Number of residues	3,166,747	109,990	344,492	151,819	209,856	142,439	356,416	248,302	288,199
Number of binding residues	590,128	22,313	65,771	31,718	39,804	29,935	64,556	48,376	49,563
Number of non-binding residues	2,576,619	87,677	278,721	120,101	170,052	112,504	291,860	199,926	238,636

Table S2: Summary of datasets for protein-peptide interaction and B-cell epitopes.

Task	B-cell epitopes					Peptide-binding sites	
Datasets	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	TR1038	TE125
Number of chains	990	573	758	926	499	1,038	125
Number of residues	434,416	142,559	259,635	307,726	138,038	248,577	30,870
Number of binding residues	51,848	19,509	27,313	38,085	15,822	13,552	1,716
Number of non-binding residues	382,568	123,050	232,322	269,641	122,216	235,025	29,154

Table S3: The information used in S2Site

Input Type	Atom	Residue
Attribute	A one-hot encoded vector $atom_i \in \mathbb{R}^{1\times 12}$ according to its bonding with C, CH, CH_2, CH_3, Cπ (aromatic carbon), O, OH, N, NH, NH_2, S, or SH	A 3B ESM-2 encoded vector $res_i \in \mathbb{R}^{1\times 2560}$
Index	The residue index that the atom belongs to	The corresponding sequence index that starts from 0 in continuous ascending order
Point Cloud	The atom's 3D coordinates	C_α atoms' 3D coordinates and the corresponding sidechain center of mass
Triplets of Neighborhood Indices	The position indices of the current atom and its two selected neighbor atoms in the point cloud	The position indices of current and previous C_α atoms as well as the corresponding sidechain center of mass in the point cloud

Table S4: Comparing S2Site against ScanNet and PeSTo* in predicting protein-binding site on the 417 structures shared among the test sets of the three methods. The evaluation is done using the median ROC AUC, AUPRC, and MCC. In ScanNet and PeSTo, the residue is labeled as a binding site if it is within 5 Å from another protein while S2Site is tested on test sets where the distance is within 4 Å instead.

Dataset	Dataset Size	ROC AUC			AUPRC			MCC		
		ScanNet	PeSTo	S2Site	ScanNet	PeSTo	S2Site	ScanNet	PeSTo	S2Site
Test 70%	98	0.893	0.912	**0.935**	0.782	0.794	**0.850**	0.560	0.593	**0.620**
Test Homology	133	0.887	0.906	**0.930**	0.691	0.753	**0.802**	0.494	0.576	**0.577**
Test Topology	87	0.921	**0.971**	0.963	0.796	**0.900**	0.852	0.579	**0.769**	0.699
Test None	99	0.850	0.842	**0.865**	0.512	**0.587**	0.582	0.363	**0.461**	0.331
Test All	417	0.897	0.929	**0.931**	0.720	0.797	**0.800**	0.510	**0.636**	0.577

* The result of ScanNet and PeSTo are taken from the PeSTo paper.

S.6 Supplementary Figures

Figure S2 to S9 visualize the changes in the two-dimensional residue-level proteins' attribute distribution with TSNE before and after the neighborhood feature extraction.

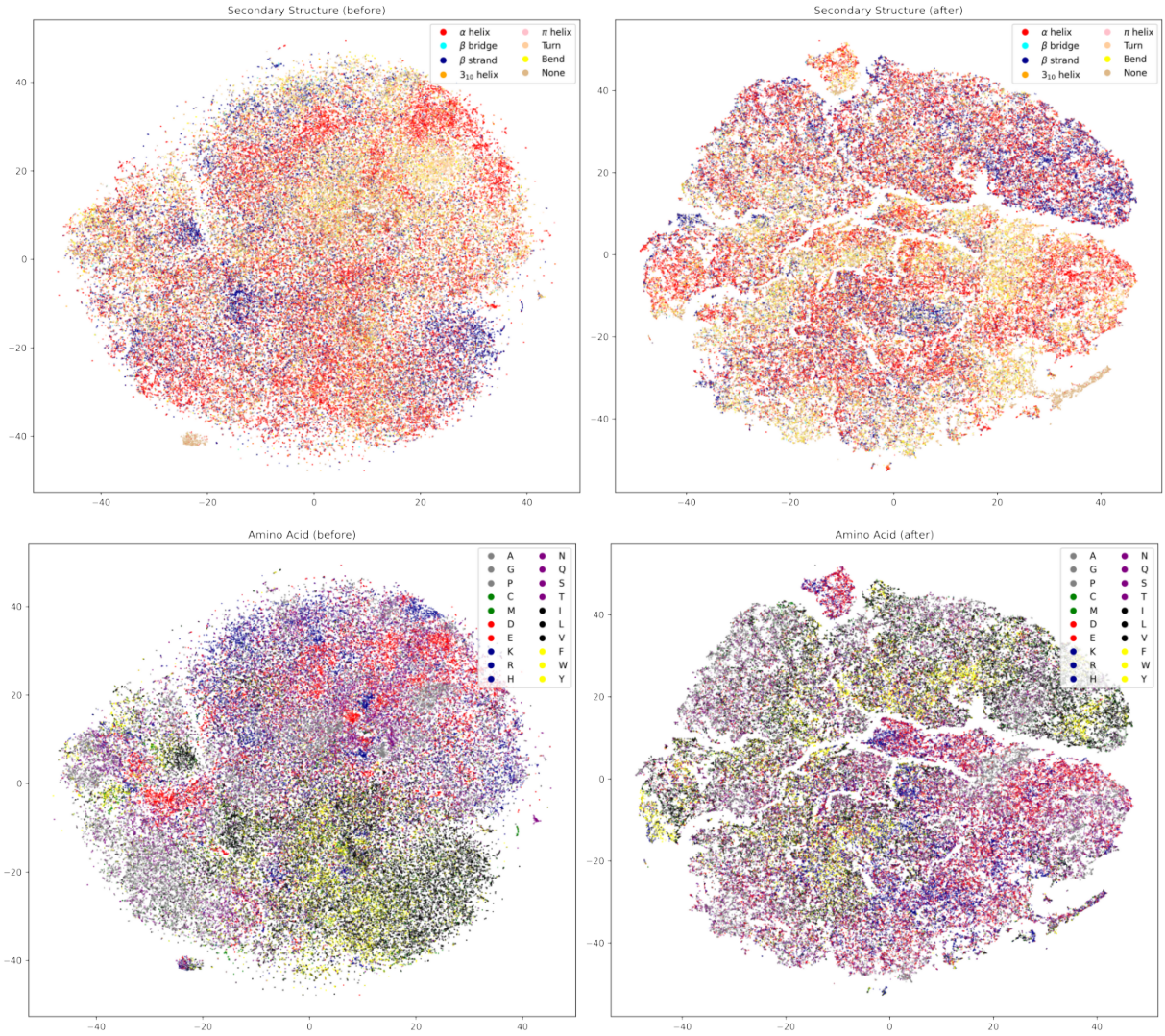

Figure S2: Illustration of the changes in the secondary structure and amino acid attribute distribution of Test 70% before and after the neighborhood feature extraction.

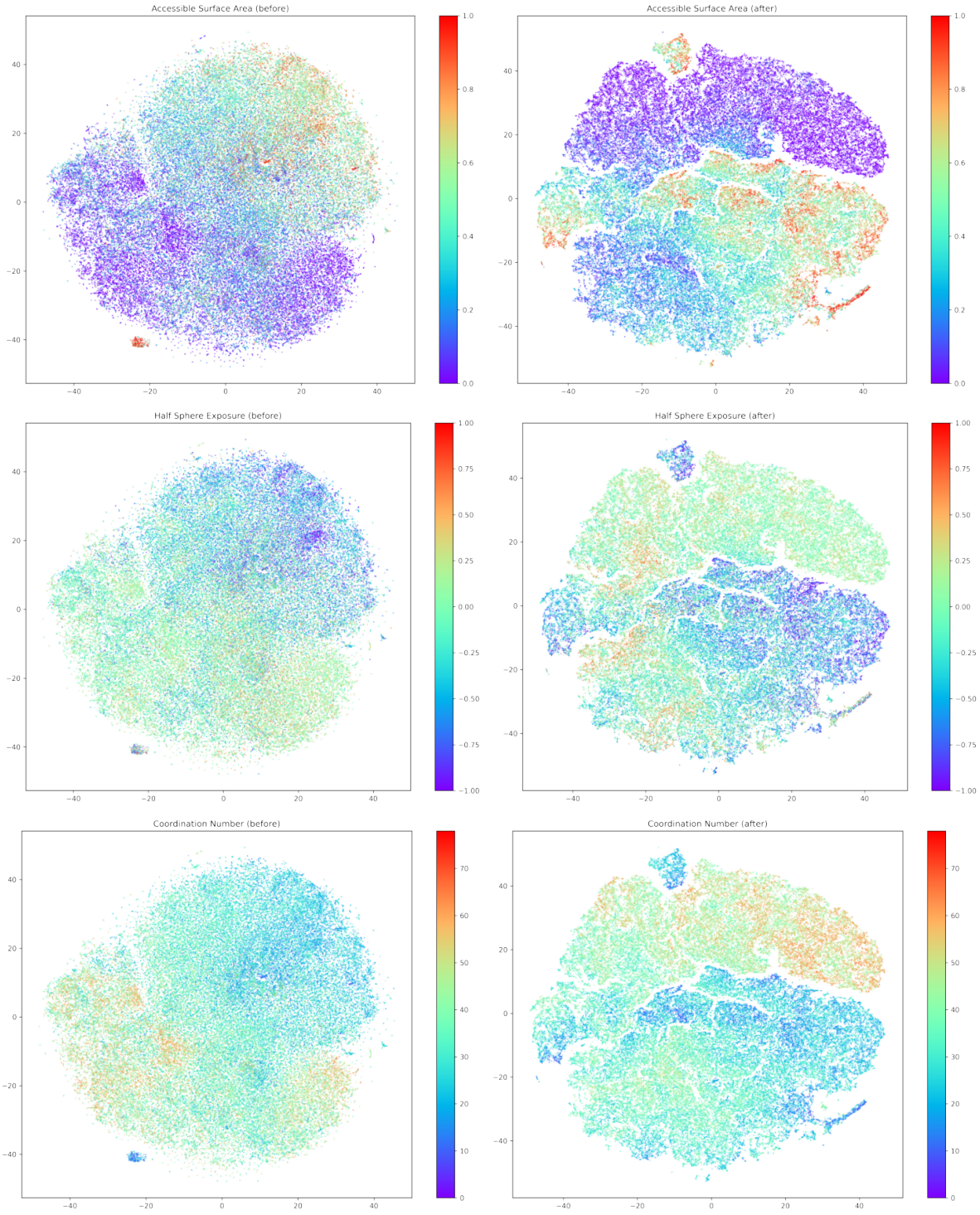

Figure S3: Illustration of the result for the accessible surface area, half sphere exposure, and coordination number attribute distribution of Test 70% before and after the neighborhood feature extraction.

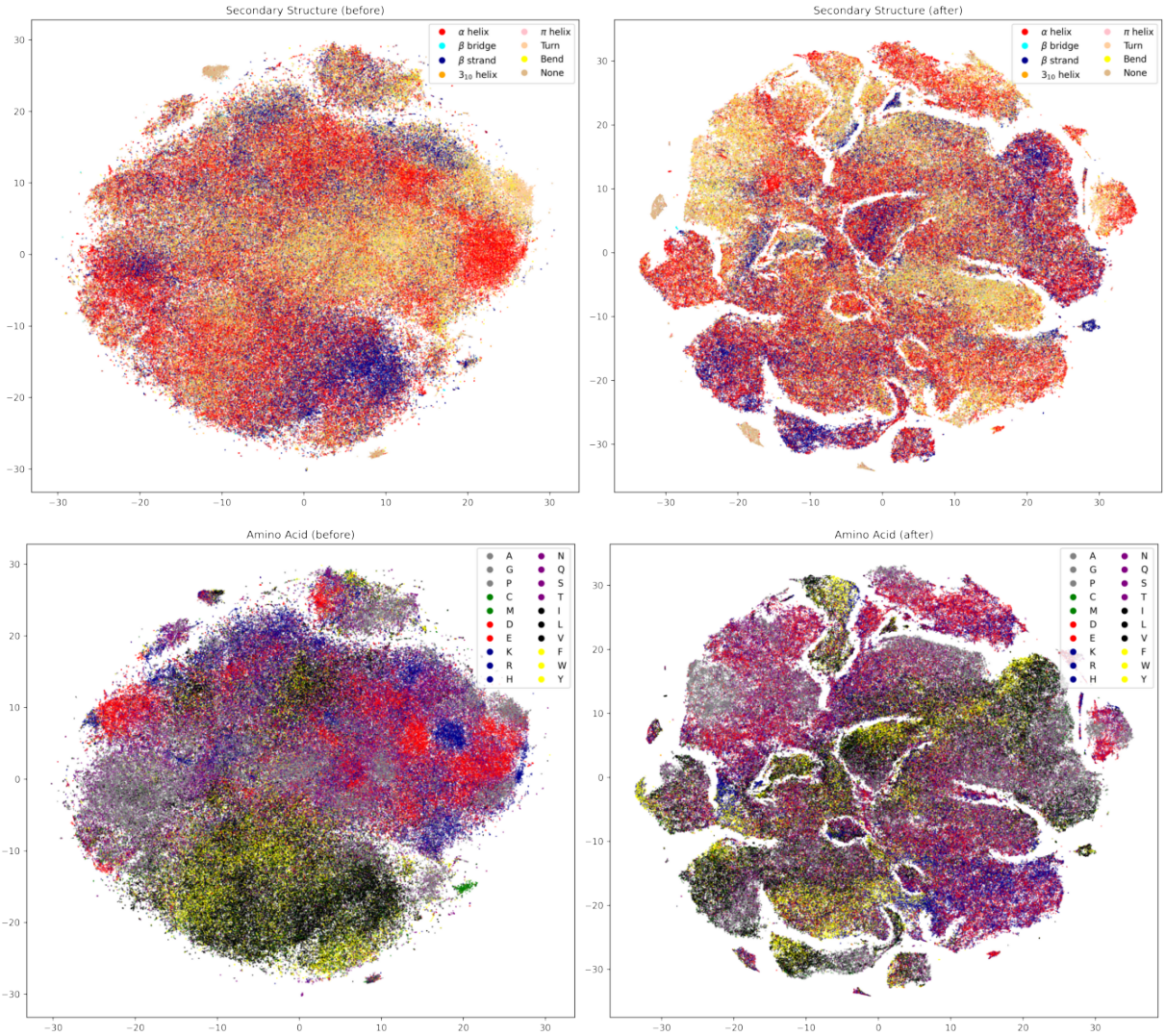

Figure S4: Illustration of the changes in the secondary structure and amino acid attribute distribution of Test Homology before and after the neighborhood feature extraction.

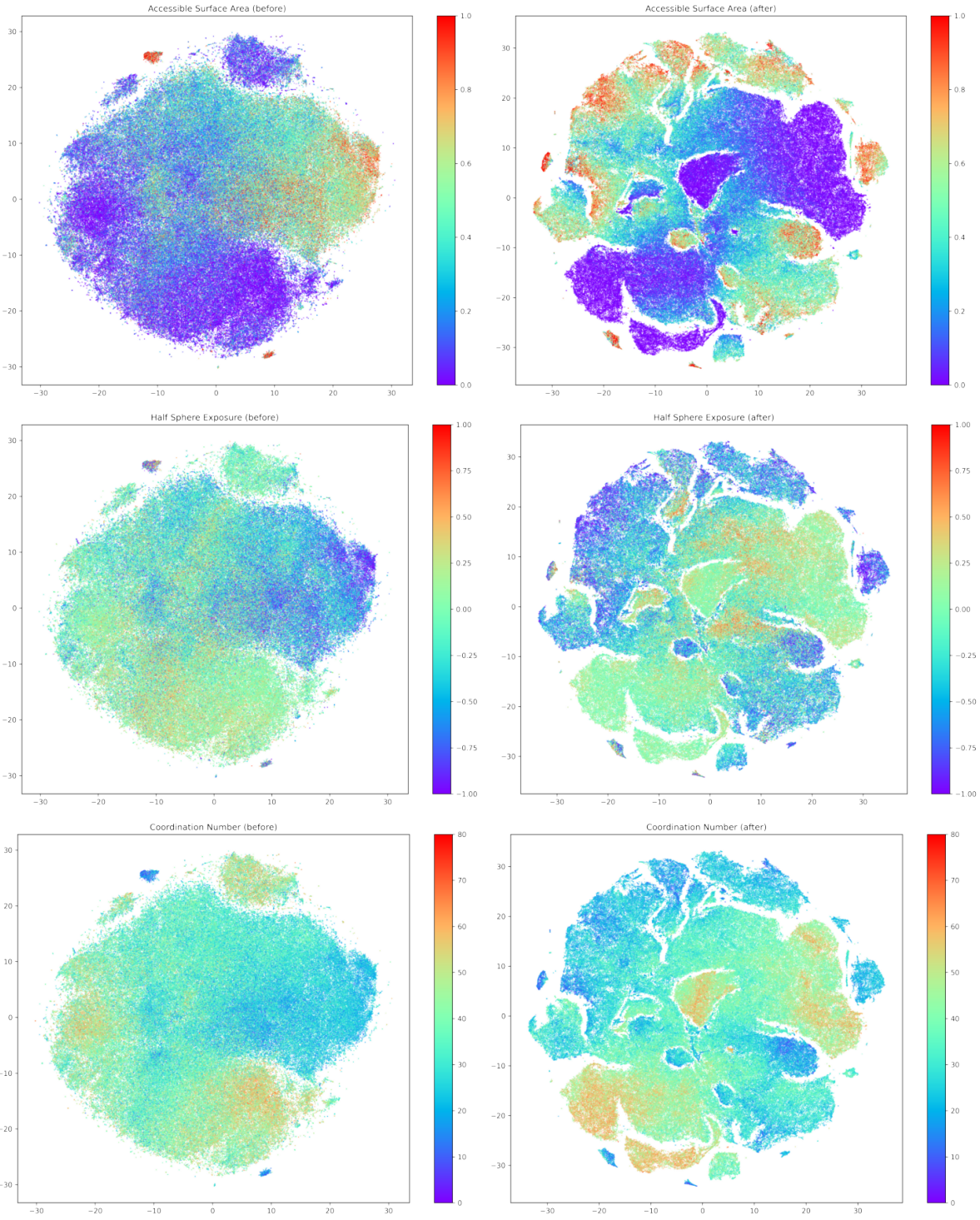

Figure S5: Illustration of the changes in the accessible surface area, half sphere exposure, and coordination number attribute distribution of Test Homology before and after the neighborhood feature extraction.

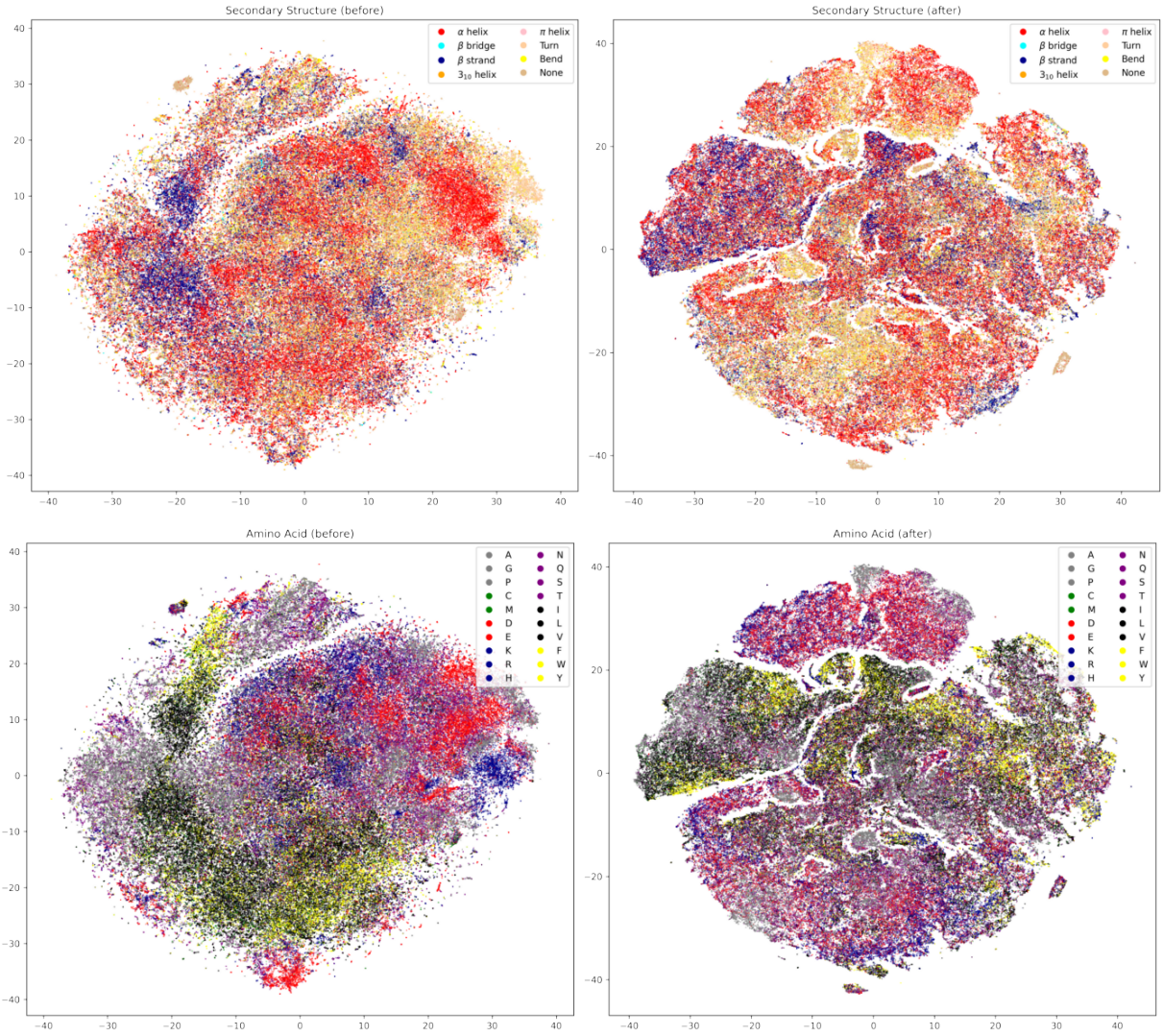

Figure S6: Illustration of the changes in the secondary structure and amino acid attribute distribution of Test Topology before and after the neighborhood feature extraction.

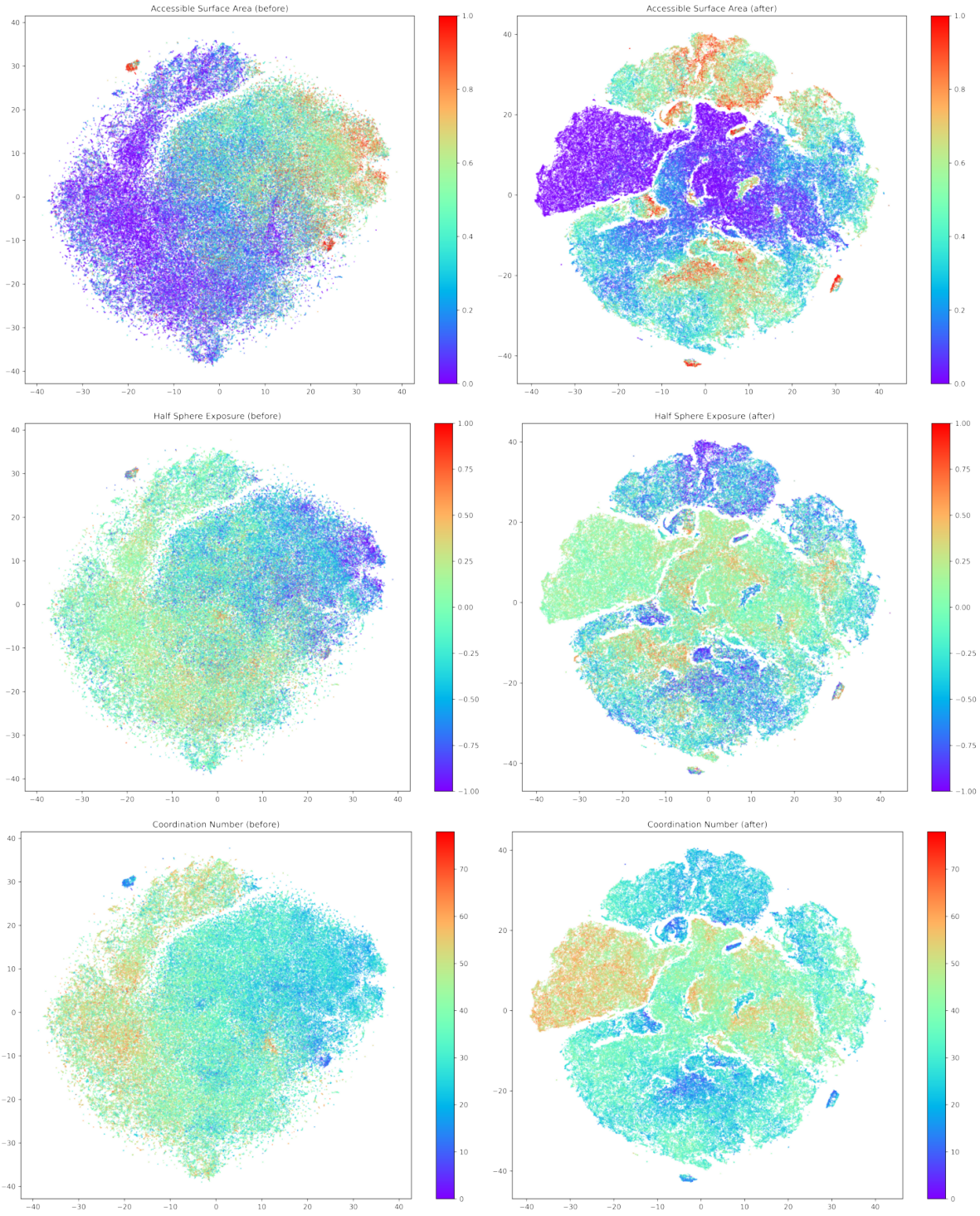

Figure S7: Illustration of the changes in the accessible surface area, half sphere exposure, and coordination number attribute distribution of Test Topology before and after the neighborhood feature extraction.

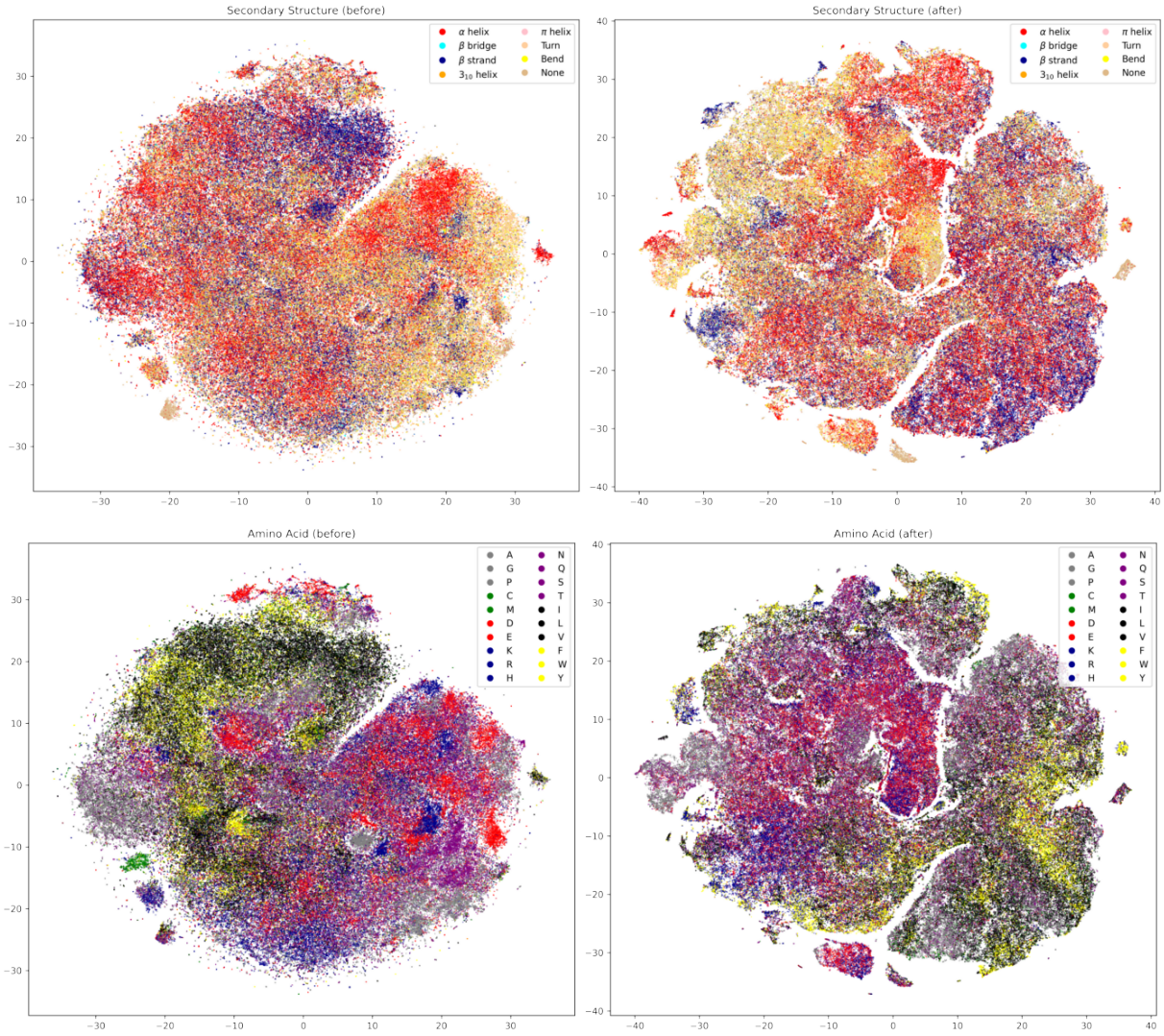

Figure S8: Illustration of the changes in the secondary structure and amino acid attribute distribution of Test None before and after the neighborhood feature extraction.

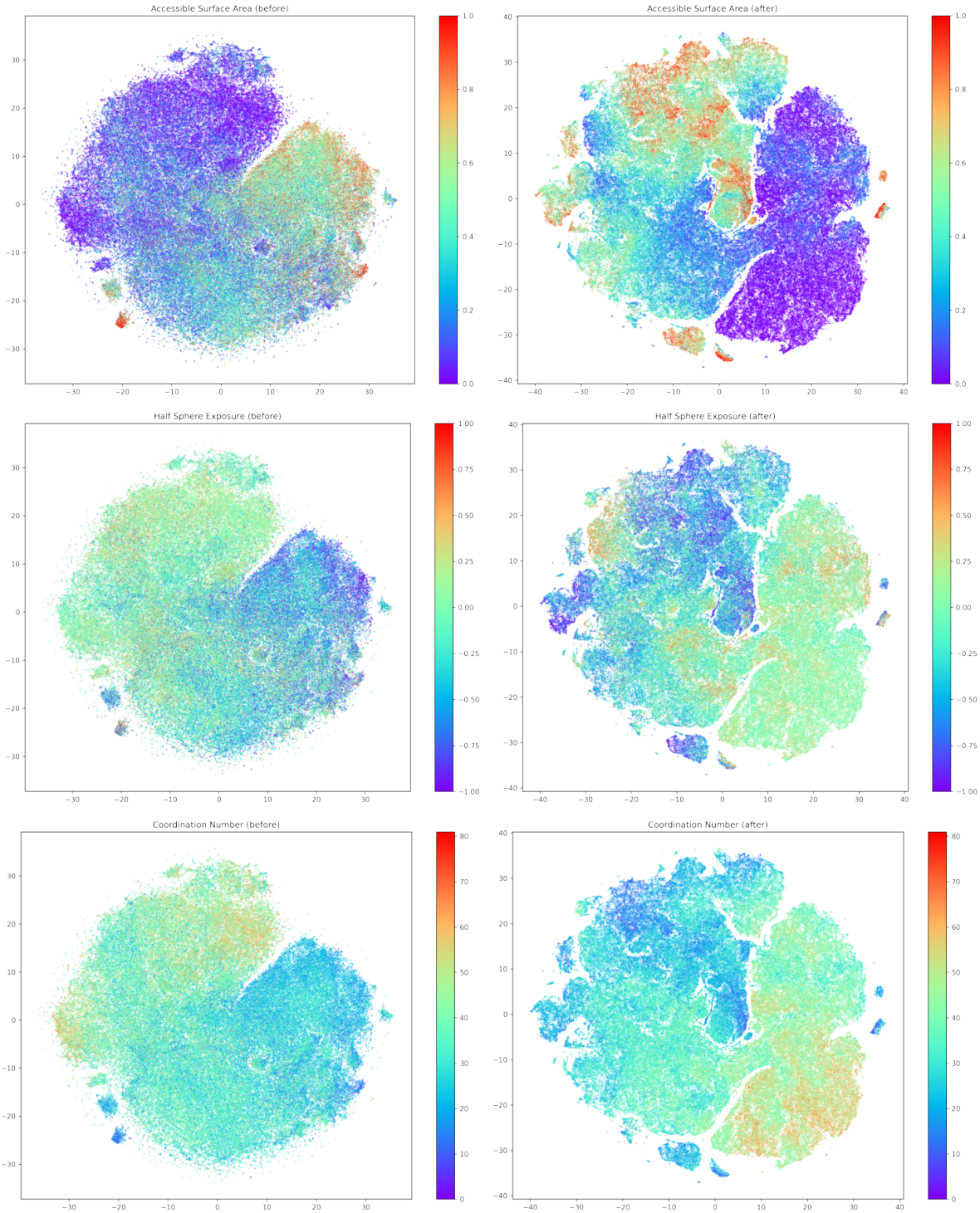

Figure S9: Illustration of the changes in the accessible surface area, half sphere exposure, and coordination number attribute distribution of Test None before and after the neighborhood feature extraction.