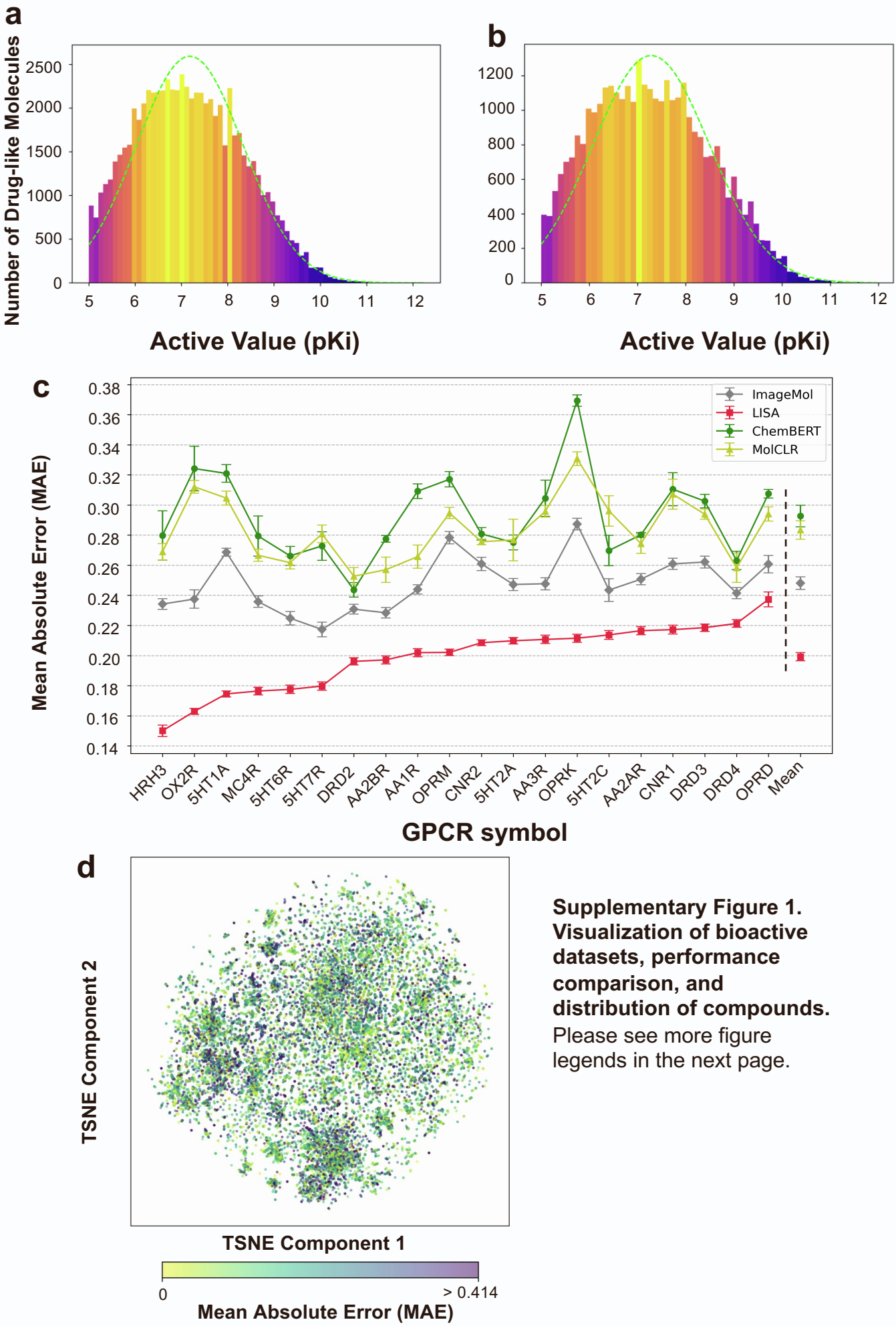
Cell Reports Methods, Volume *4*

Supplemental information

A deep learning framework combining molecular image and protein structural representations identifies candidate drugs for pain

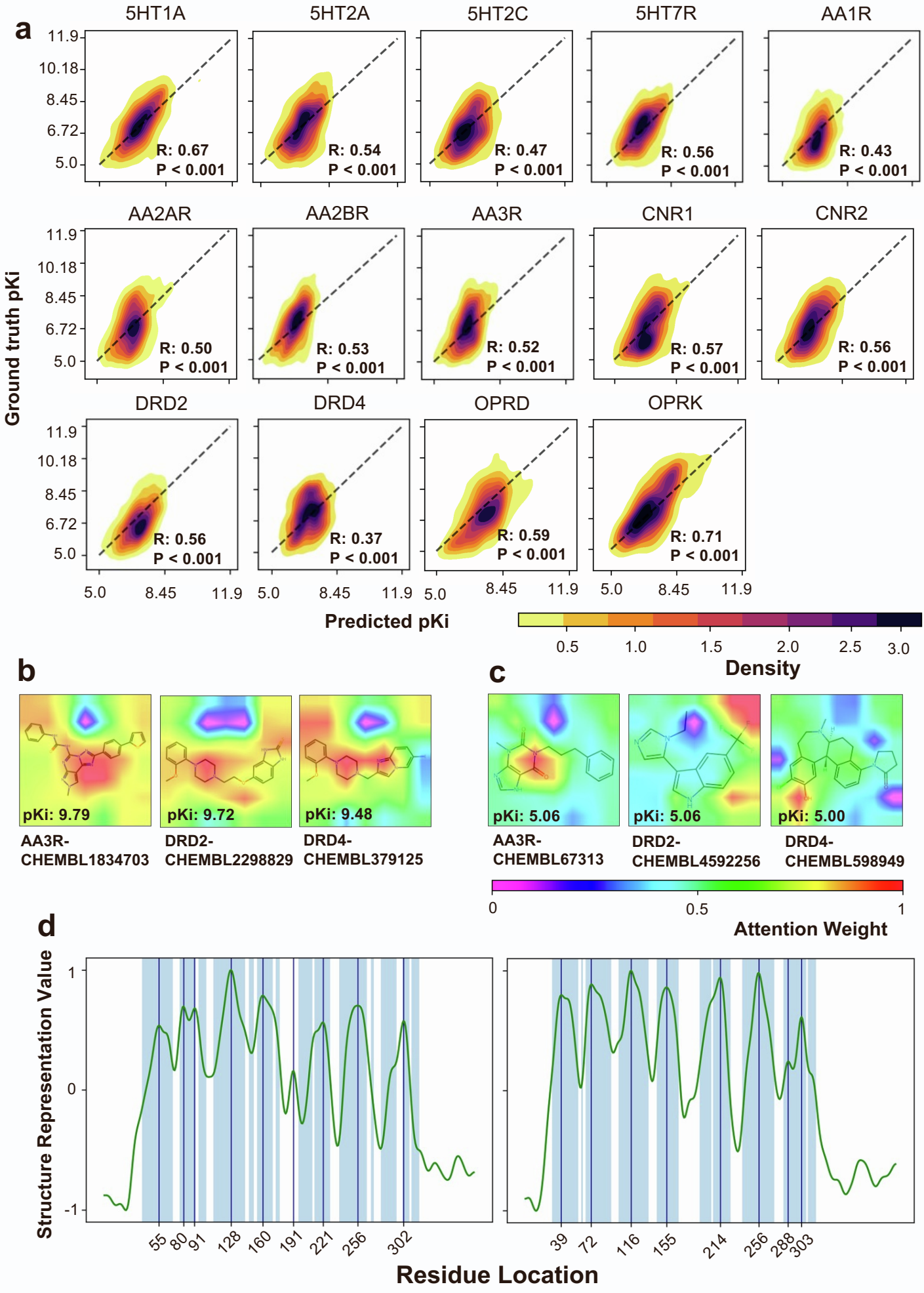
Yuxin Yang, Yunguang Qiu, Jianying Hu, Michal Rosen-Zvi, Qiang Guan, and Feixiong Cheng

Supplementary Figures 1-6



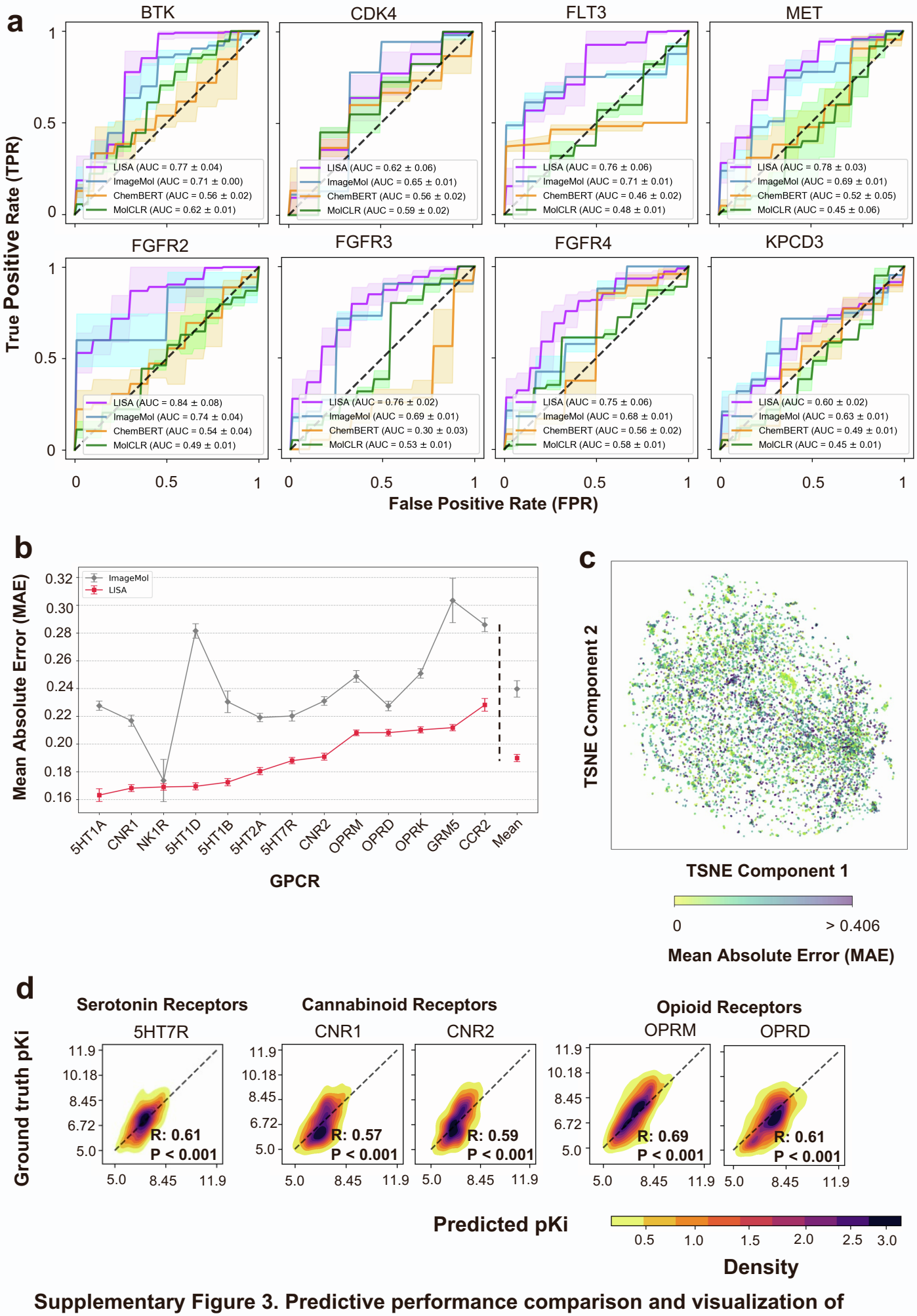
Supplementary Figure 1. Visualization of bioactive datasets, performance comparison, and distribution of compounds. Please see more figure legends in the next page.

Supplementary Figure 1. Visualization of bioactive datasets, performance comparison, and distribution of compounds. a, Distribution of active value (pKi) of compounds in the top-20 GPCR dataset. 76.6% of all compounds have an activity value between 6 and 9, and the mean activity value of the dataset is 7.18. **b,** Distribution of pKi of compounds in the pain-related GPCR dataset. 74.9% of all compounds have an activity value between 6 and 9, and the mean activity value of the dataset is 7.28. Deeper color indicates small number of compounds in the corresponding active value ranges, and lighter color indicates the opposite. We use a Gaussian distribution curve (green dashed line) to fit the distribution of data. **c,** Mean absolute error (MAE) between the original ImageMol model (gray line), CHEM-BERT (green line), MolCLR (yellow line), and the proposed LISA-CPI model (red line) on the 20 GPCR targets in the top-20 GPCR dataset and the mean values of the MAEs. Error bars indicate the uncertainty of both models measured by standard deviation. **d,** Distribution of compounds in the test set of the top-20 GPCR dataset visualized using TSNE. Predictive performance of each single compound is also visualized with colors. Warmer colors (colors close to yellow) indicate lower mean absolute error (MAE) or better performance. Cooler colors (colors close to purple) indicate higher mean absolute error (MAE) or worse performance. Visualization of MAE is capped at 0.414, which 90% of compounds in the test set are predicted with MAE lower than this value.



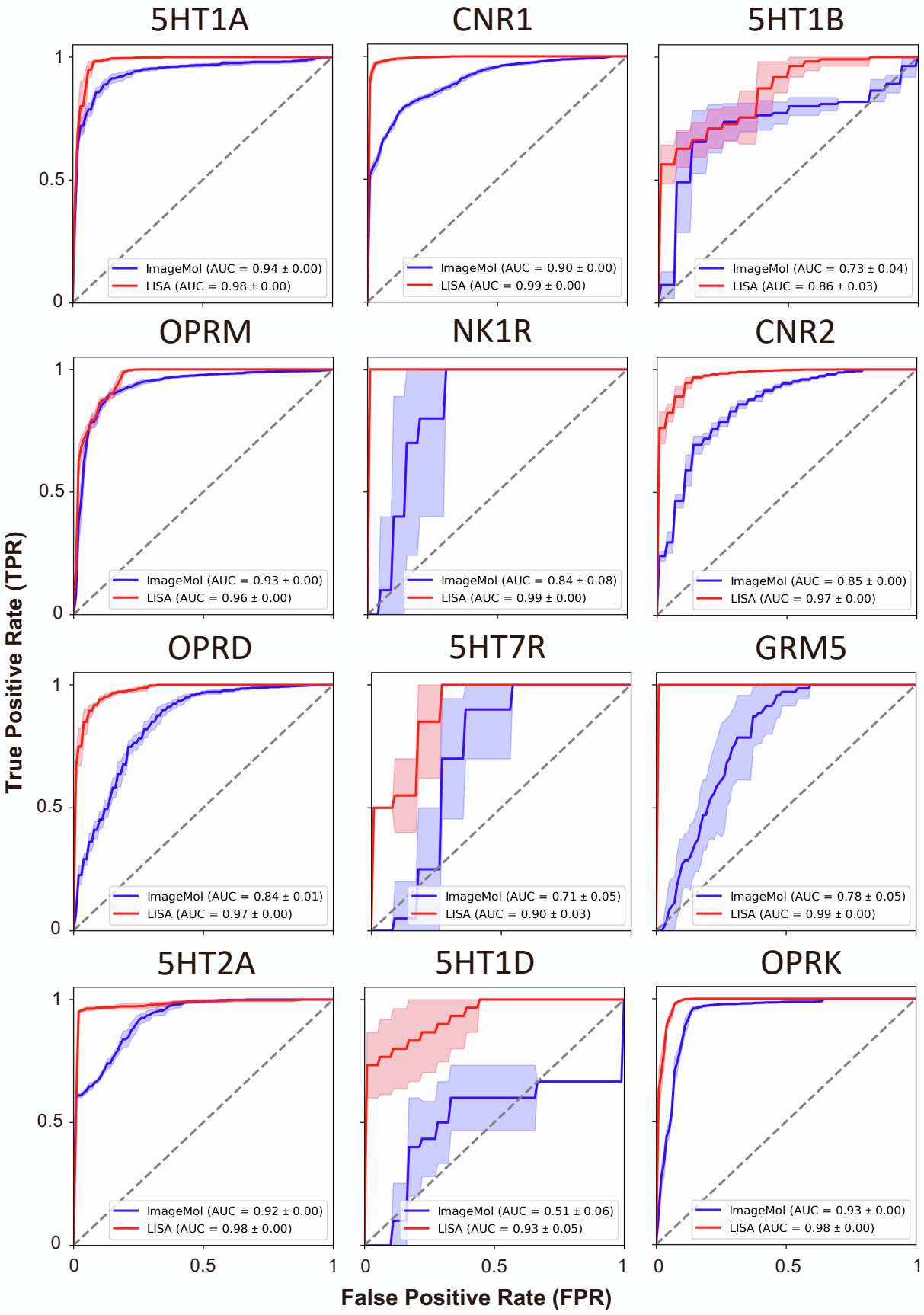
Supplementary Figure 2. Predictive performance, representative attention heatmaps, and structure representations. Please more detailed figure legends in the next page.

Supplementary Figure 2. Predictive performance, representative attention heatmaps, and structure representations. a, Predictive performance of our proposed LISA-CPI on the rest of 14 of top-20 GPCR targets. Predicted pKi and actual pKi of each compound for each GPCR target are contour plotted with points density. Pearson's correlation coefficient (R) and p values are labeled for each GPCR dataset. **b,** Attention patterns and binding structures of compounds with high active value (pKi > 8). **c,** Attention patterns and binding structures of compounds with low active value (pKi < 6). Warmer colors (colors close to red) indicate higher attention and cooler colors (colors close to purple) indicate lower attention. **d,** Visualization of structure representation of CCR2 (left). Visualization of structure representation of NK1R (right). Structure representations are scaled to the range of [-1,1] using min-max normalization. Peaks of structure representations are marked using blue vertical lines. Light blue areas indicate the positions of helical regions which contain key information about structures and functions of proteins.

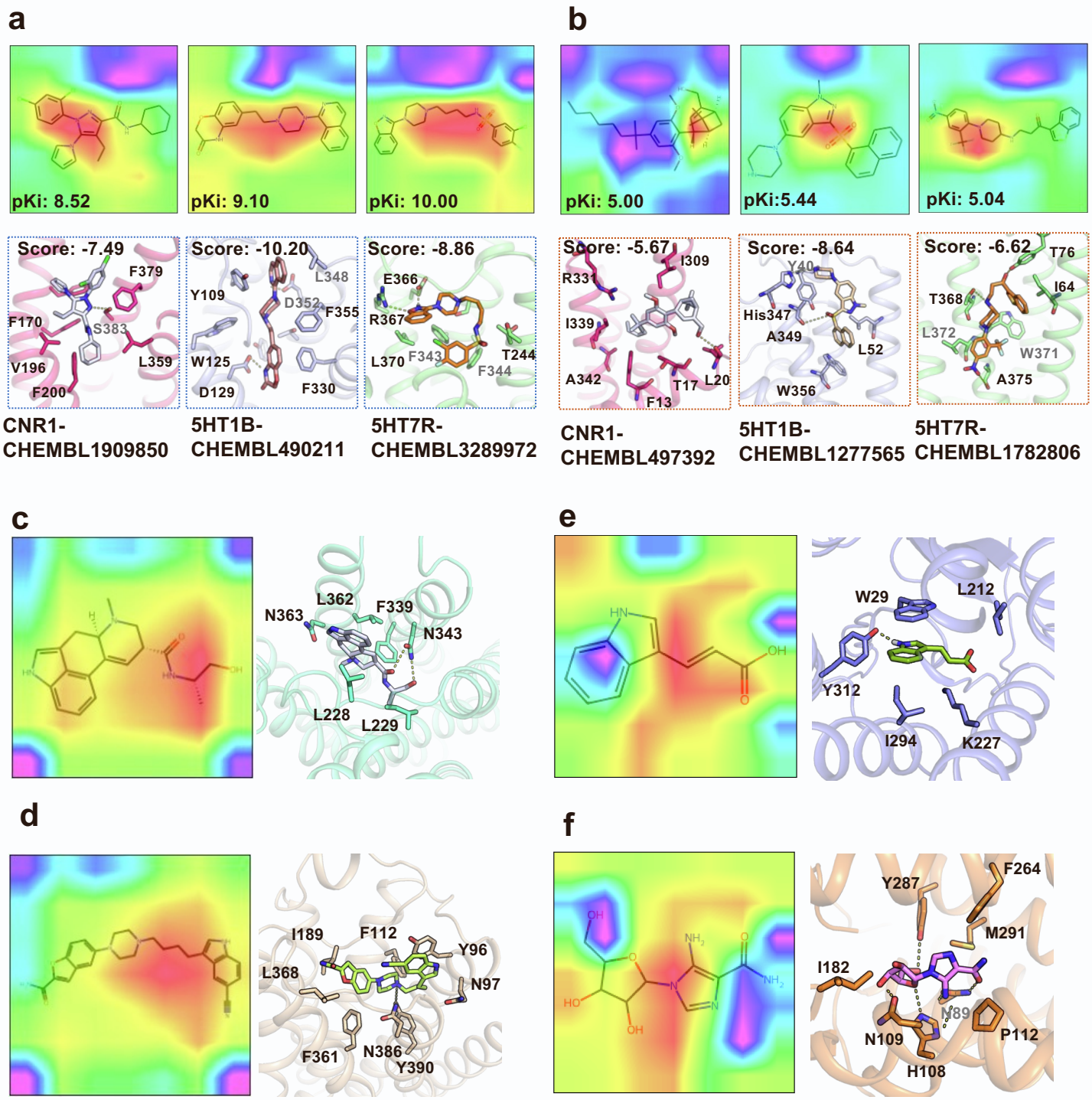


Supplementary Figure 3. Predictive performance comparison and visualization of compounds distribution. Please more detailed figure legends in the next page.

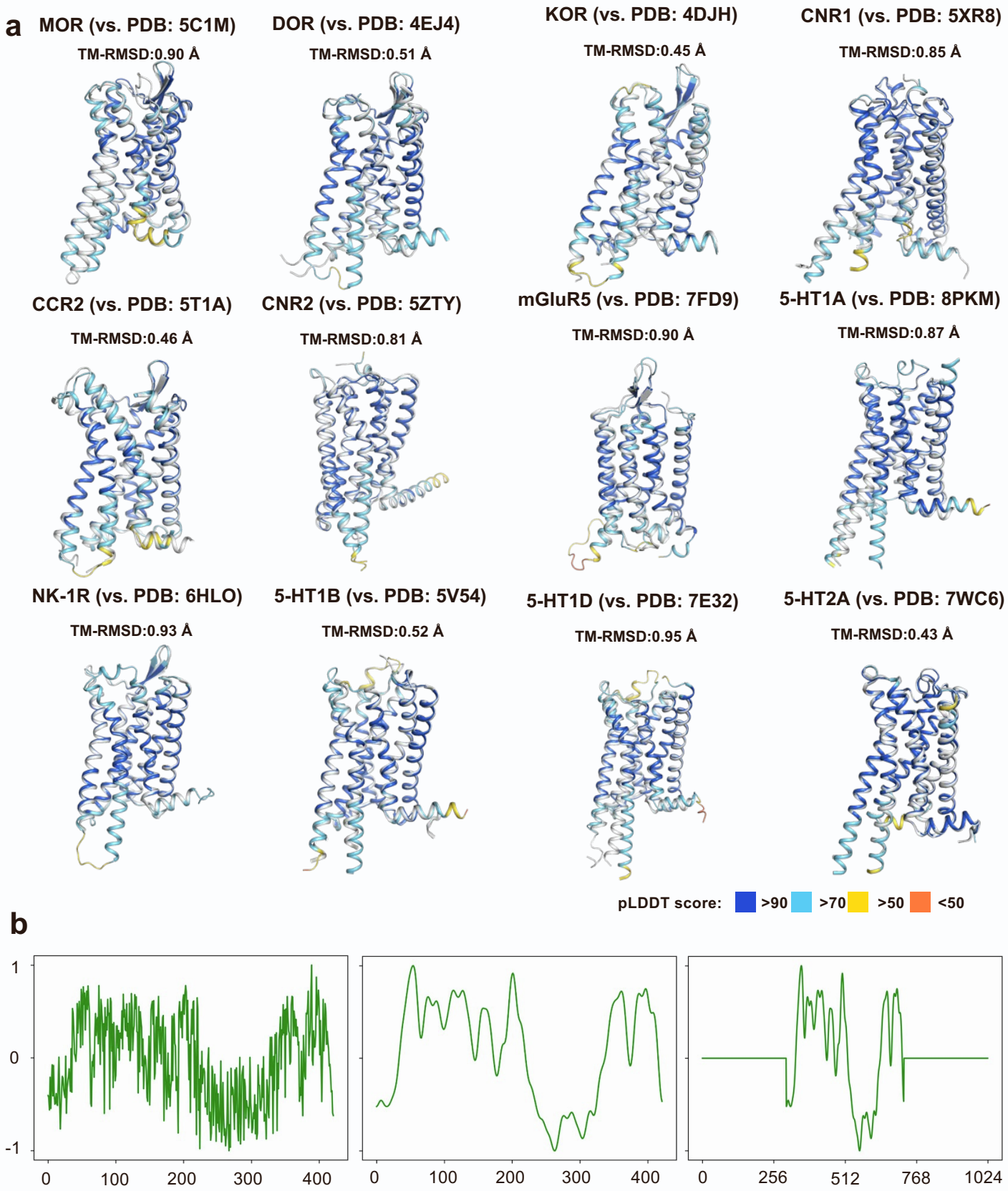
Supplementary Figure 3. Predictive performance comparison and visualization of compounds distribution. a, Receiver Operating Characterisitc (ROC) curves showcasing the predictive performance of LISA-CPI and three baseline models (ImageMol, CHEMBERT, and MolCLR) on rest of the 8 Kinase datasets. Solid lines and shades represent the mean and one standard deviation of ROC curves obtained from 10-fold cross validation. **b,** Mean absolute error (MAE) between the original ImageMol model (gray line) and the proposed LISA-CPI model (red line) on the 13 GPCR targets in the pain-related GPCR dataset and the mean values of the MAEs. Error bars indicate the uncertainty of both models measured by standard deviation. **c,** Predictive performance of our proposed LISA-CPI on the rest of 5 pain-related GPCR targets categorized by different receptors. Predicted pKi and actual pKi of each compound for each GPCR target are contour plotted with points density. Pearson's correlation coefficient (R) and p values are labeled for each GPCR dataset. **d,** Distribution of compounds in the test set of the 13 pain-related GPCR dataset visualized using TSNE. Predictive performance of each single compound is also visualized with colors. Warmer colors (colors close to yellow) indicate lower mean absolute error (MAE) or better performance. Cooler colors (colors close to purple) indicate higher mean absolute error (MAE) or worse performance. Visualization of MAE is capped at 0.406, which 90% of compounds in the test set are predicted with MAE lower than this value.



Supplementary Figure 4. Predictive performance of LISA-CPI and ImageMol on the agonist-antagonist datasets. Solid lines and shades represent the mean and one standard deviation of ROC curves obtained from 10-fold cross validation.



Supplementary Figure 5. Representative heatmaps and putative binding modes targeting pain-related GPCRs. a, Heatmaps of attention levels on ligand images with high activity values, where pKi of these compounds are greater than 8 (first row) and putative binding modes of these molecules with their corresponding receptors (second row). **b,** Heatmaps of attention levels on ligand images with low activity values, where pKi of these compounds are smaller than 6 (first row) and putative binding structures of these molecules with their corresponding receptors (second row). **c and d,** Drug repurposing predictions targeting pain-related GPCRs. **c,** Left: attention pattern of Ergometrine on 5HT2A (antagonist), right: the putative binding structure of Ergometrine on 5HT2A. **d,** Left: attention pattern of Vilazodone on 5HT1A (agonist), right: the putative binding structure of Vilazodone and 5HT1A. **e and f,** Gut-microbiota derived metabolite repurposing predictions targeting pain-related GPCRs. **e,** Left: attention pattern of Indoleacrylic Acid on OPRK (agonist), right: the putative binding structure of Indoleacrylic Acid on OPRK. **f,** Left: attention pattern of AICAR on NK1R (antagonist), right: the putative binding structure of AICAR on NK1R. Warmer color indicates higher attention, and cooler color indicates lower attention.



Supplementary Figure 6. Structure comparison and visualization of processing of structure representations. a, Structural comparison between AlphaFold2 and crystal structure for pain related GPCRs. Root Mean Square Deviation of transmembrane region (TM-RMSD) between AlphaFold2 and crystal structures are calculated by PyMOL. AlphaFold2 models are depicted in pLDDT score. Crystal structures are depicted in color gray. Regions with pLDDT score > 70 indicates confident (blue and cyan). **b,** Visualization of processing steps of structure representations of 5HT1A. Left: Original highly noisy structure representation of 5HT1A. Middle: Smoothed structure representation of 5HT1A. Right: Zero padded structure representation of 5HT1A. Structure representations are first scaled to the range of [-1,1] using min-max normalization, followed by Gaussian smoothing and zero padding.