Supplementary Information

Supplementary Algorithm S1 Ray casting algorithm

Algorithm 1 Ray Casting Algorithm

Input: ligand 3D PH4 point $p \in \mathbb{R}^{n\times 3}$, protein vertices $\mathcal{V} \in \mathbb{R}^{n\times 3}$, $\varepsilon = 10^{-9}$

1: Construct convex hull $\mathcal{H} = \text{ConvexHull}(\mathcal{V})$

2: Extract triangles of polyhedron $\boldsymbol{t} = \mathcal{H}.\text{simplices}$

3: Set ray origin point $o \leftarrow [1.0, 0.0, 0.0]$

4: Initialize intersection count c

5: **for** t in $\boldsymbol{t}$ **do**

6: Extract triangle vertices $(v_0, v_1, v_2) = \mathcal{V}[t]$

7: Compute edges $\boldsymbol{e_1} = v_1 - v_0$, $\boldsymbol{e_2} = v_2 - v_0$

8: Compute determinant $d = \boldsymbol{e_1} \cdot (\boldsymbol{o} \times \boldsymbol{e_2})$ (Eq. 1)

9: **if** $|d| < \varepsilon$ **then**

10: continue to next triangle

11: Compute barycentric coordinates $u = \frac{1}{d}[(\boldsymbol{o} - v_0) \cdot (\boldsymbol{o} \times \boldsymbol{e_2})]$ (Eq. **2**)

12: **if** $u < 0$ or $u > 1$ **then**

13: continue to next triangle

14: Compute barycentric coordinates $v = \frac{1}{d}\{\boldsymbol{o} \cdot [(\boldsymbol{o} - v_0) \times \boldsymbol{e_1}]\}$ (Eq. **3**)

15: **if** $v < 0$ or $u + v > 1$ **then**

16: continute to next triangle

17: Compute ray intersection distance $t = \frac{1}{d}\{\boldsymbol{e_2} \cdot [(\boldsymbol{o} - v_0) \times \boldsymbol{e_1}]\}$ (Eq. **4**)

18: **if** $t > \varepsilon$ **then**

19: $c \leftarrow c + 1$

Output: boolean(c mod 2 == 0)

Supplementary Algorithm S2 Molecular voxelization and featurization algorithm

Algorithm 2 Molecular Voxelization and Featurization Algorithm

Input: ligand 3D atomic structure $\mathcal{M}_{i=1}^{N} \in \mathbb{R}^{x\times y\times z}$, protein 3D atomic structure $\mathcal{P} \in \mathbb{R}^{x\times y\times z}$, 3D voxelized box boundary $\mathcal{B} \in [[x_{\min}, x_{\max}], [y_{\min}, y_{\max}], [z_{\min}, z_{\max}]]$

1: Obtain PH4 feature (f) representation $\mathcal{U}\{(x, y, z), f\}_{i=1}^{N} = \text{GetFeaturesForMol}(\mathcal{M}_{i=1}^{N}), f \in \{0, \ldots, 5\}$

2: Convert coordinates $\mathcal{U}(x, y, z)_{i=1}^{N}$ into 3D grid (g) points $\mathcal{G}_{i=1}^{N} \in \mathbb{R}^{x_g\times y_g\times z_g}$, where $x_g \in [x_{\min}, x_{\max}], y_g \in [y_{\min}, y_{\max}], z_g \in [z_{\min}, z_{\max}]$

3: Embed protein pocket $\mathcal{P}$ into 3D grid vertices $\mathcal{V} \in \mathbb{R}^{x_v\times y_v\times z_v}$, where $\mathcal{V} = \text{ReadPocket}(\mathcal{P})$

4: **for** i in $1, \ldots, N$ **do**

5: **if** $\mathcal{G}$ is outside convex hull formed by $\mathcal{V}$ **then**

6: $\mathcal{U}(f_i) = -1$

7: **end if**

8: **end for**

9: Obtain feature array $\mathcal{U}\{G, F\}_{i=1}^{N} \in \mathbb{R}^{G\times F}$, where $\mathcal{U}_{i=1}^{N} \leftarrow \text{Flatten}(\mathcal{U}\{(x_g, y_g, z_g), f\ != -1\}_{i=1}^{N})$

Output: flatten one-hot 3D feature matrix $\mathcal{F} \in \mathbb{R}^{G\times F\times M}$

Supplementary Algorithm S3 Matrix simplification algorithm

Algorithm 3 Matrix Simplification Algorithm

Input: flatten one-hot 3D feature matrix $\mathcal{F} \in \mathbb{R}^{G\times F\times M}$

1: Reverse feature matrix $\mathcal{F} \leftarrow \sim\mathcal{F}$

2: Element-wise multiplication for all compounds $\mathcal{D} = \prod_{i=1}^{N} \mathcal{F}$

3: Let attention index $\mathcal{Z} := \{(i, j) \mid \mathcal{D}[i][j] = 0\}, \mathcal{Z} \in \mathbb{R}^{G\times 2}$

4: Initialize simplified matrix $\mathcal{T} \in \mathbb{R}^{M\times G}$

5: **for** i in $1, \ldots, N$ **do**

6: **for** j in $1, \ldots, G$ **do**

7: $\mathcal{T}_{ij} = \mathcal{F}_i[\mathcal{Z}^G, \mathcal{Z}^F]$

Output: 2D attention index $\mathcal{Z}$, 2D trainable tensor $\mathcal{T}$

Supplementary Algorithm S4 Training and weighting procedures of Ph3DG

Algorithm 4 Training and Weighting Procedures of Ph3DG

Input: 2D trainable tensor $\mathcal{T}$, 2D attention index $\mathcal{Z}$, 3D voxelized box boundary $\mathcal{B} \in [[x_{\min}, x_{\max}], [y_{\min}, y_{\max}], [z_{\min}, z_{\max}]]$

1: Initialize weight tensor w

2: **while** $\mathcal{L}_\theta$ not converge **do**

3: $f_\theta, y_\theta \leftarrow \text{SplitDataset}(\mathcal{T})$

4: $\hat{y} = \text{pooling}(f_\theta \odot w_\theta)$ (Eq. 6)

5: $\mathcal{L} = \text{MSE}(\hat{y}, y_\theta)$ between predicted label $\hat{y}$ and true label y_θ (Eq. 7)

6: Update w_θ by minimizing $\mathcal{L}$

7: Obtain training output weight tensor $W \in \mathbb{R}^{G}$ by model.state_dict()

8: Obtain weight-grid mapping $\mathcal{M} \in \mathbb{R}^{G\times 3}$ containing grid G, feature type F, and weight W via $\mathcal{Z} \in \mathbb{R}^{G\times 2}$

9: Obtain 3D grid coordinate $grid\_x, grid\_y, grid\_z \leftarrow \text{Unravel}(\mathcal{M}^G)$

Output: weighted grid $((G, F, W, x_g, y_g, z_g))$

Supplementary Table S1. Benchmarking performance of different baseline models.

Task	Metric	Ph3DG	EquiScore	PLANET	Glide	Phase	Autodock-VIna
Training	AUROC	0.928 ± 0.132	**0.955 ± 0.082**	0.784 ± 0.173	0.164 ± 0.256	0.765 ± 0.221	0.660 ± 0.177
	AUPRC	**0.997 ± 0.003**	0.656 ± 0.208	0.758 ± 0.189	0.352 ± 0.306	0.789 ± 0.268	0.657 ± 0.163
	Success(1%)	**0.738 ± 0.139**	0.503 ± 0.336	0.592 ± 0.207	0.351 ± 0.354	0.500 ± 0.500	0.567 ± 0.324
	Success(5%)	0.719 ± 0.167	0.537 ± 0.313	**0.745 ± 0.103**	0.261 ± 0.331	0.500 ± 0.500	0.545 ± 0.334
	Success(10%)	0.687 ± 0.160	0.541 ± 0.307	**0.797 ± 0.158**	0.229 ± 0.295	0.498 ± 0.498	0.552 ± 0.336
	BEDROC(α=20)	0.015 ± 0.002	**0.087 ± 0.045**	0.022 ± 0.012	0.008 ± 0.009	0.012 ± 0.006	0.009 ± 0.002
Screening	Recall	0.254 ± 0.111	0.044 ± 0.059	**0.348 ± 0.368**	0.169 ± 0.116	0.113 ± 0.116	0.061 ± 0.112
	Enrichment	4.69 ± 4.42	0.453 ± 0.621	**4.84 ± 4.98**	3.09 ± 2.55	1.75 ± 1.77	0.907 ± 1.26
	Ranking↓	0.549 ± 0.257	0.651 ± 0.209	**0.349 ± 0.289**	0.361 ± 0.309	0.445 ± 0.311	0.382 ± 0.261
	Recalled target	**8**	3	5	6	5	6
	Recalled conformer	44	11	**51**	39	17	15

[a] All results are in percentage except for Enrichment factor. Best results are highlighted in **bold**.

Supplementary Table S2. Detailed model settings for Ph3DG-MLP variants.

Data set	Ph3DG	Ph3DG (w/o EV)
Number of training molecules	2765	2773
Number of datapoints	13735	13780
Feature dimensions	5916	11876
Hyperparameters		
Fold of cross validation	5	
Number of epochs	10000	
Batch size	16	
Learning rate	0.001	
Early stopping patience	500	
Data set split type	StratifiedKFold (shuffle)	

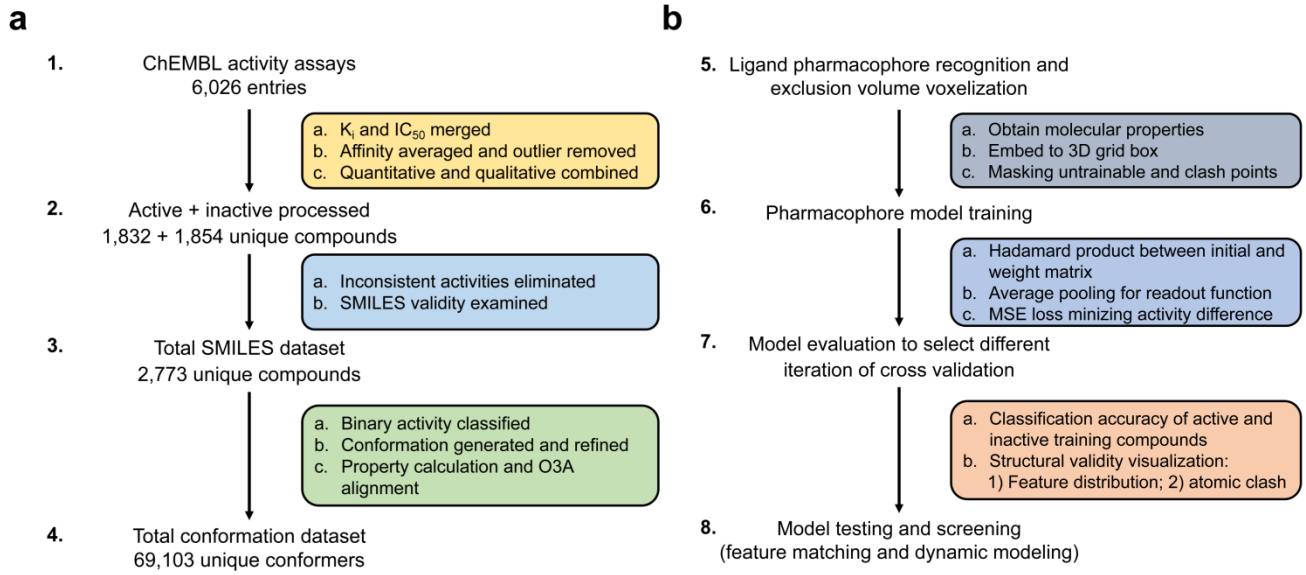
[a] Hyperparameters listed are specific for NK1R data set. For other targets parameter settings are not specifically adjust with different feature dimensions and number of data points.

[b] Ph3DG (w/o EV) indicates Ph3DG model without exclusion volume constraints.

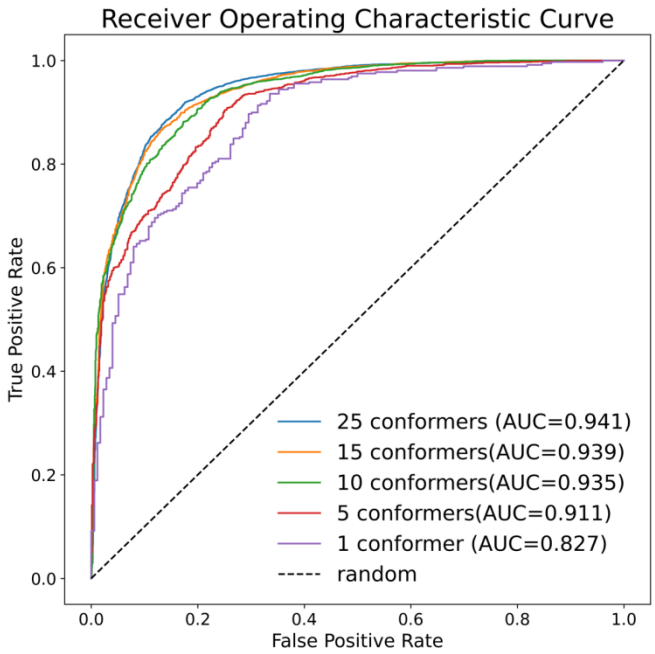
Supplementary Table S3. Detailed parameter settings for Ph3DG-Diff.

Hyperparameters	Ph3DG-Diff
Number of fully-connected layer	3
Number of neurons for each layer	5917-5916-1
Timesteps	100
Range of noise (β)	0.001 ~ 0.2
Fold of cross validation	5
Number of epochs	1000
Batch size	32
Learning rate	10^{-4}
Early stopping patience	50

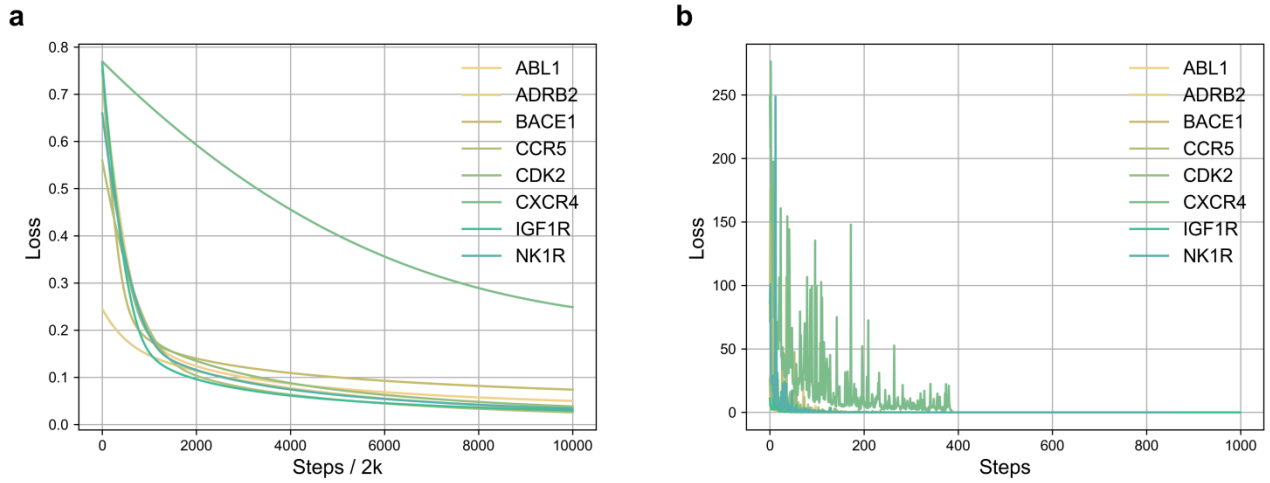
[a] Number of neurons listed are specific for NK1R data set. For other targets parameter settings are fixed.



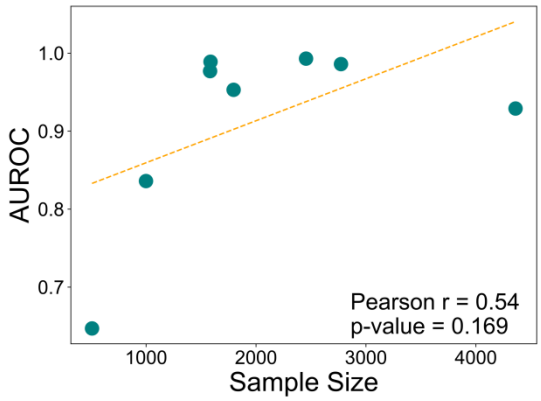
Supplementary Figure S1. Data preparation and preprocessing. **a**.Data cleaning procedures for NK1R as an exampler. Step 1. Target-specific ligand binding activity data are searched and filtered through the ChEMBL database. Step 2. Compounds are classified into 'active' and 'inactive' categories based on activity labels. Step 3. 3D conformational representations are generated by RDKit and embedded in PH4 feature numeric form. **b**. Data preprocessing procedures. Step 5. Excluded and untrainable grids are recognized and eliminated to prepare input array. Step 6. Model training with initial input array and random weight matrix. Step 7. Model valuation considering predicted PH4 score and PH4 distribution.



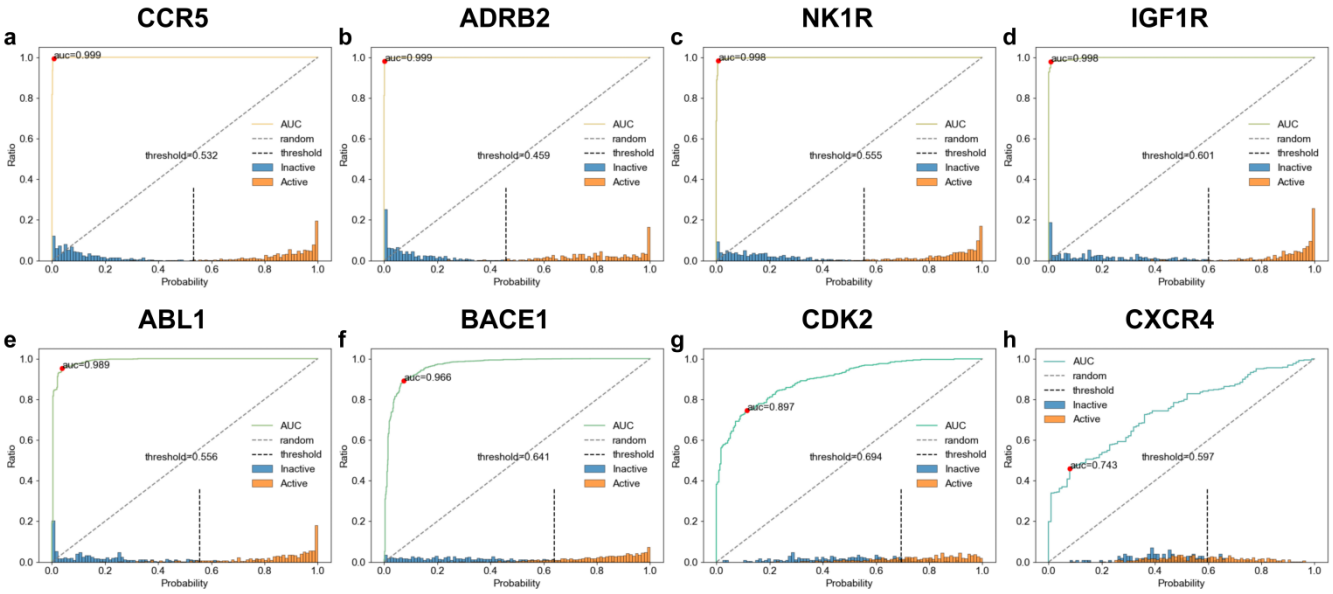
Supplementary Figure S2. Preliminary study of the optimal number of embedded conformers. AUROC classification accuracy of Ph3DG-MLP model is used to evaluate the performance of embedding different number of conformers. The more conformers per molecule embedded, the more accurate the model is.



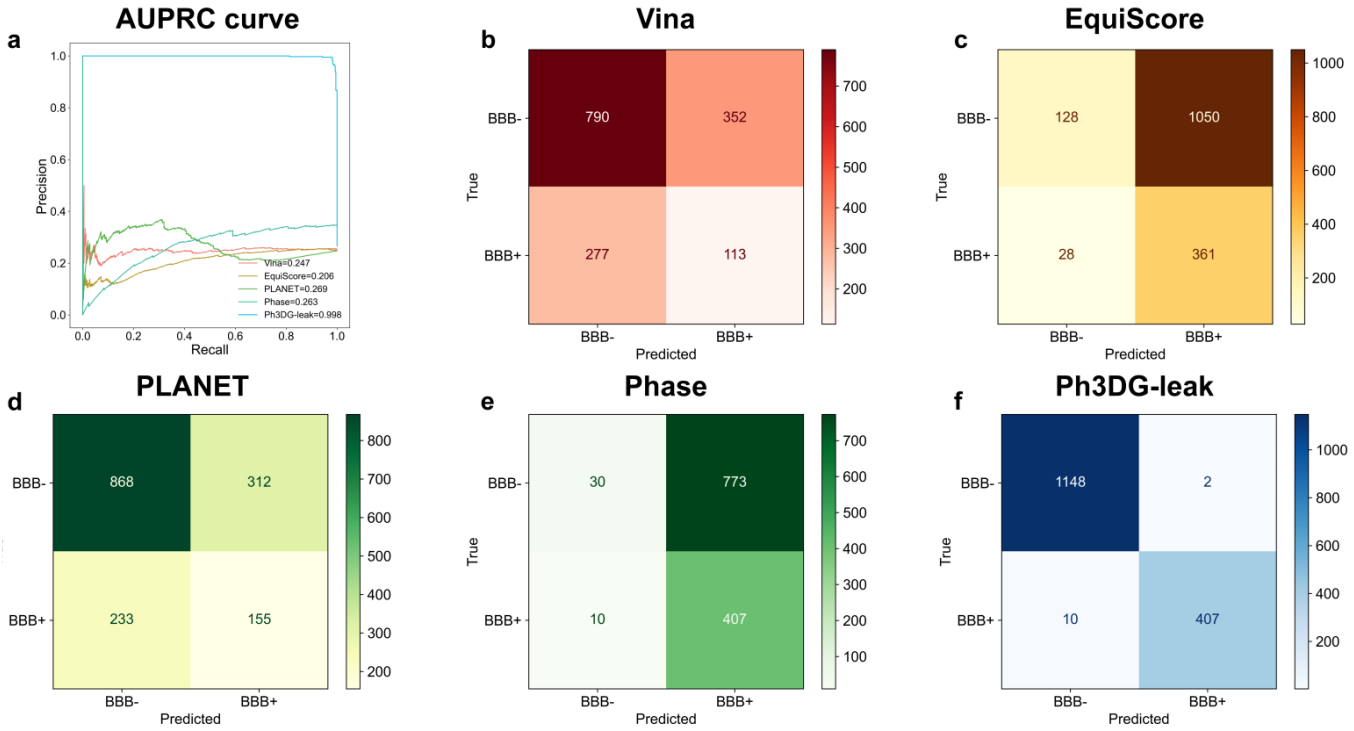
Supplementary Figure S3. The evaluation of objective function at the training stage of Ph3DG-MLP (a) and Ph3DG-Diff (b). Training loss of the first training epoch for different targets are shown. Early stop is applied to Ph3DG-Diff if MSE loss function is not decreasing within 50 epochs.



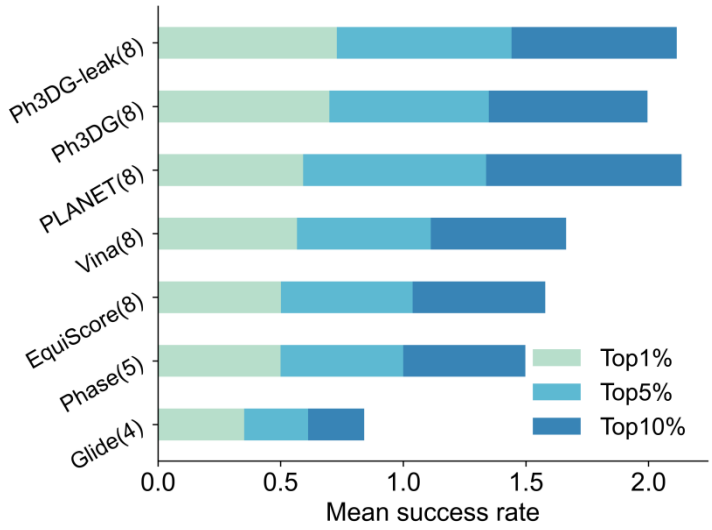
Supplementary Figure S4. Correlation between Ph3DG prediction AURCO and size of training dataset. Orange dashed line represent the fitting curve of all benchmarking systems. Detailed benchmarking targets see methods ("**Training data collection**" section).



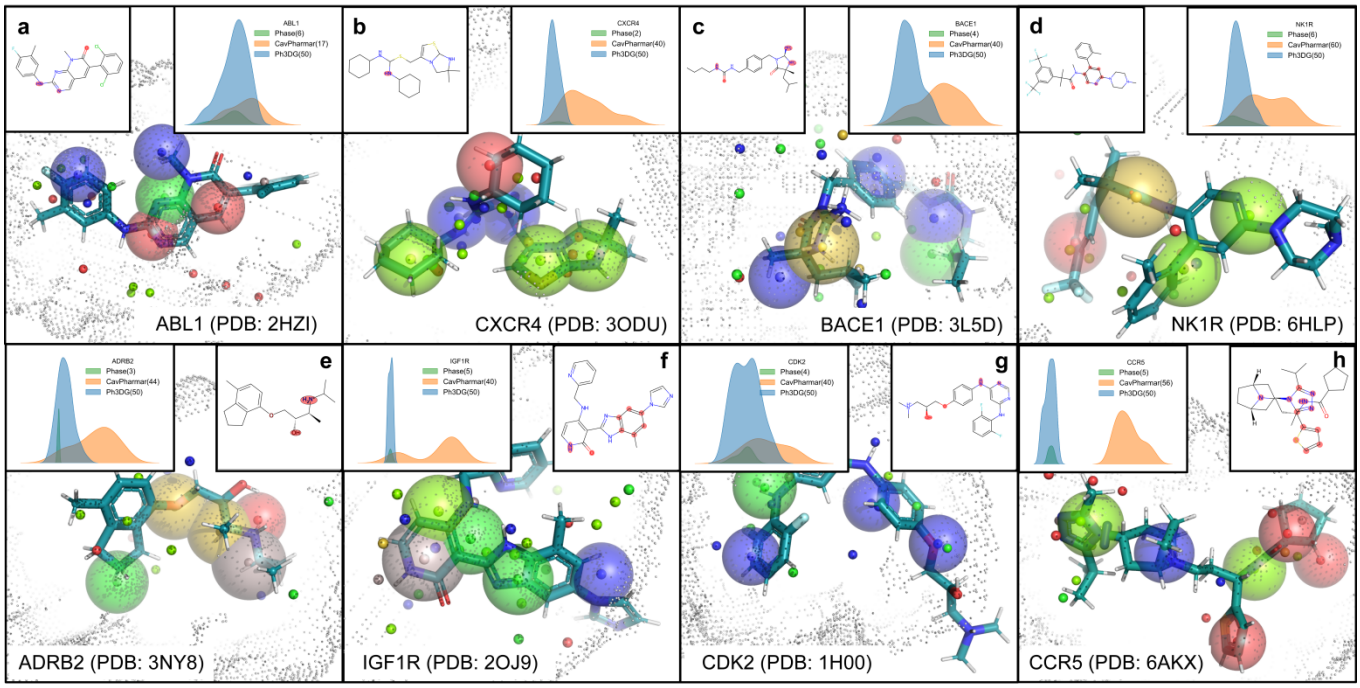
Supplementary Figure S5. Ph3DG-MLP performance of bioactivity prediction by multiple benchmarking targets. Colors of AUROC curves correspond to **Fig. 3b** in main text**.**



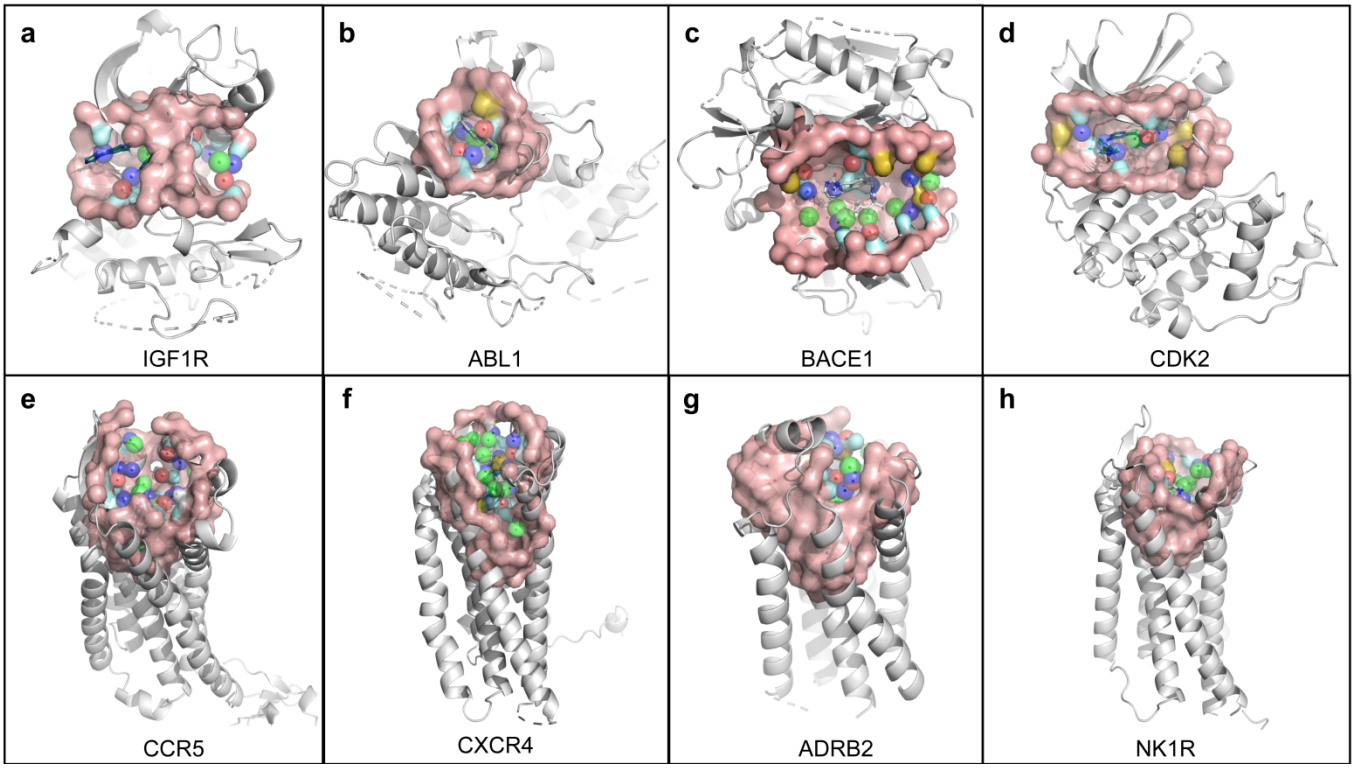
Supplementary Figure S6. Bioactivity prediction of ADRB2 target for multiple benchmarking methods.



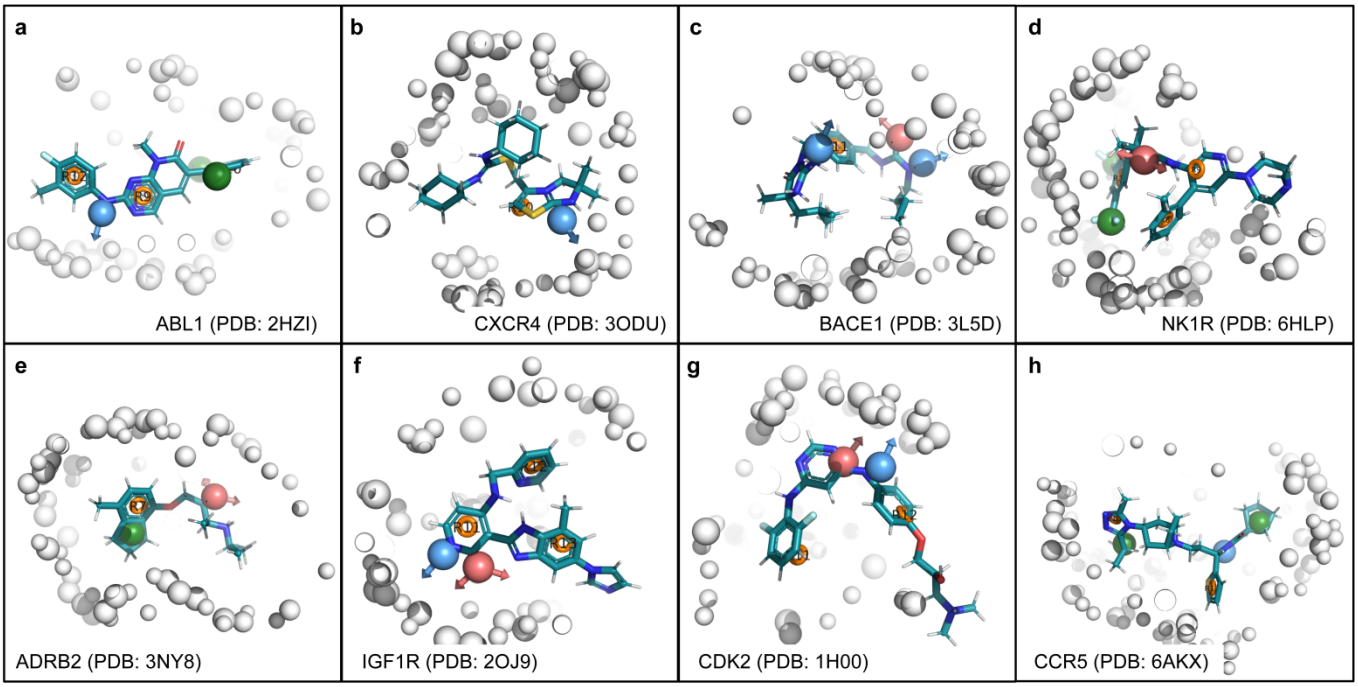
Supplementary Figure S7. Barplot of success rates compared to benchmarking methods. Results show the robust potential to retrieve positives form the top 1% active compounds of Ph3DG-MLP model. Source data are provided in Zenodo.



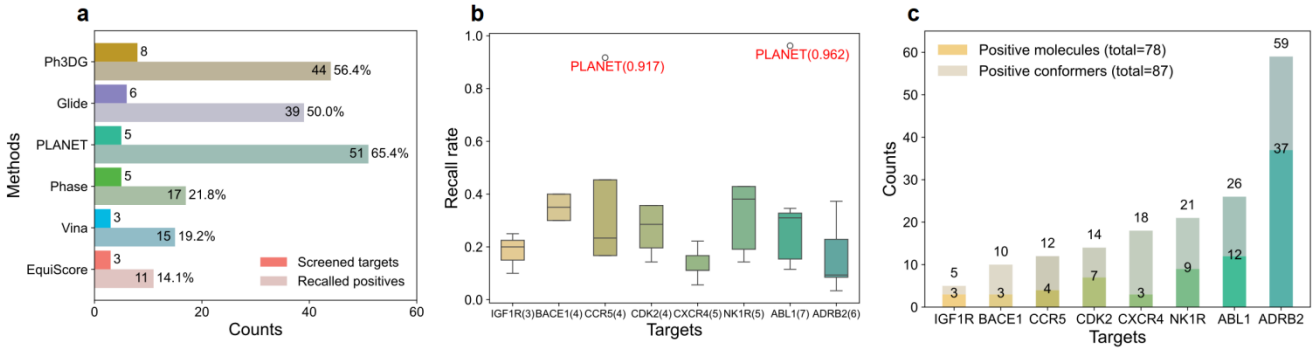
Supplementary Figure S8. Interpretability of the Ph3DG in constructing PH4 models. The essential PH4 features (top-10 features matching reference ligand shown in large transparent sphere, top-50 features shown in small sphere) and the boundary of PH4 features (shown in gray dots) captured by Ph3DG. Chemical structures of reference compounds (inactive groups highlighted) and distance distributions of PH4 features are exhibited in subfigure. PH4 distances (Ph3DG in blue, CavPharmer in orange, and Phase in green) are calculated by a specific PH4 features towards the center of mass (COM) of the reference. PH4 feature colors correspond to those shown in **Fig. 3c** in main text.



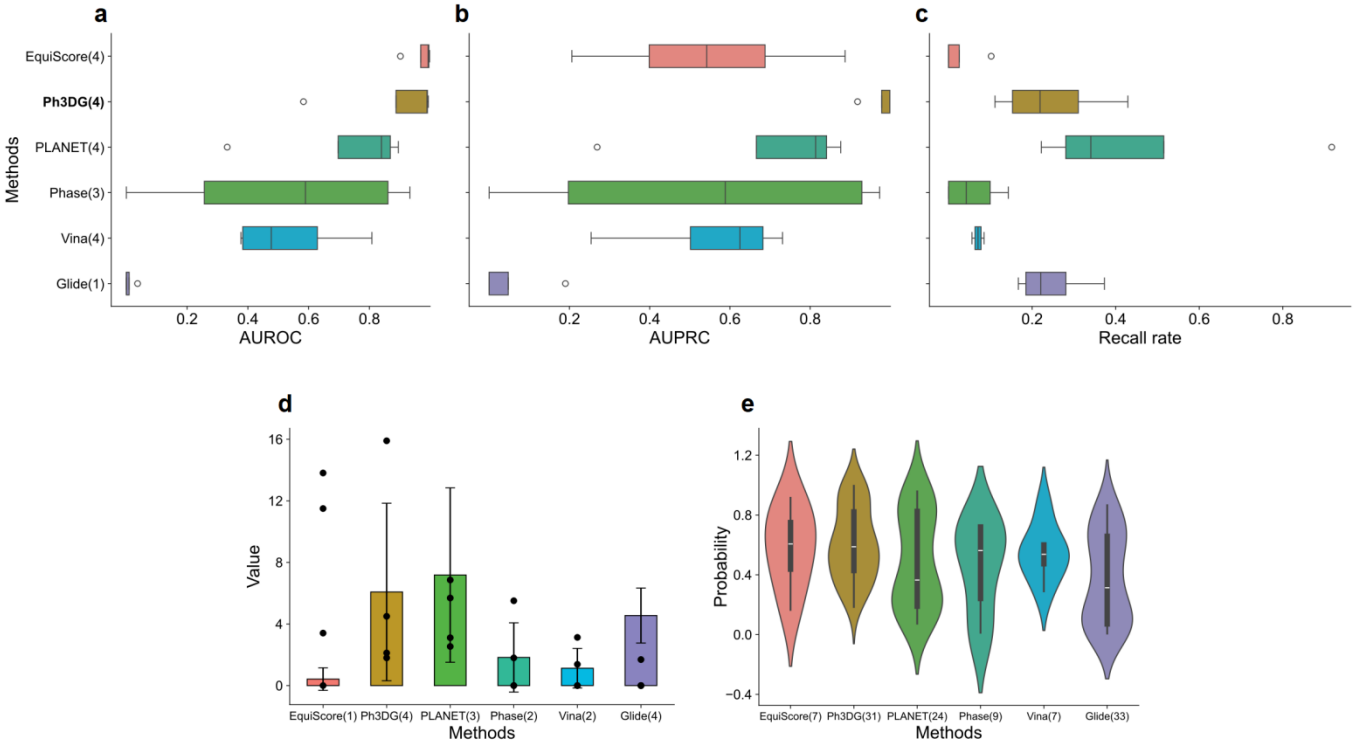
Supplementary Figure S9. Schematic representation of CavPharmer modeled PH4. Receptors are in white cartoon. Exclusion volumes are represented in pink surface, while key PH4 are indicated by spheres in different colors. Blue: HBD; red: HBA; orange: positive electrostatic charges; yellow: negative electrostatic; gray: hydrophobic; green: root of HBD/HBA. CavPharmer exhibits comparatively sparse distribution patterns for all benchmarking systems.



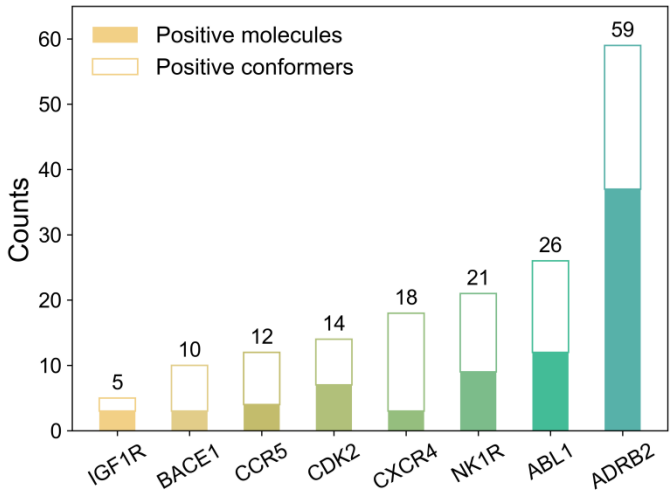
Supplementary Figure S10. Schematic representation of Phase modeled PH4 features. Ligands are represented in turquoise stickers. Exclusion volumes are represented in white spheres, while key PH4 are indicated by spheres in different colors. Light blue: HBD; pink: HBA; green: hydrophobic; orange: aromatic ring. Arrows represent the specific directions of hydrogen bonds. Phase developed pharmacophore models based on a single protein-ligand complex, thus demonstrating densely distributed PH4 as depicted in **Fig. S8**.



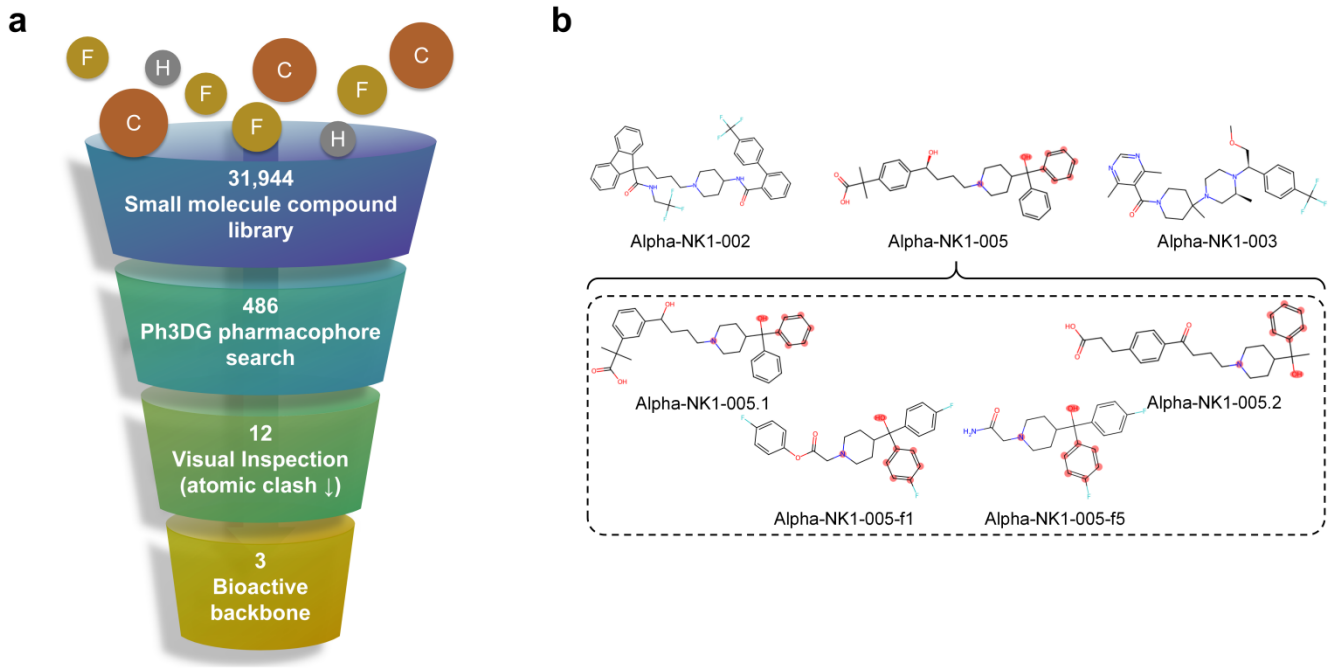
Supplementary Figure S11. Boxplot statistics of screening recalls. **a.** Counts of all prediction methods recalling benchmarked targets. Color correspond to **Fig. 2h** in main text. Percentages of the recalled molecules among all TPs in all benchmarked systems are labeled. **b.** Recall rates of benchmarking methods across all benchmarked targets. Numbers in brackets indicate the count of screening methods successfully retrieved positive compounds. PLANET prediction outliers with its values are labeled in red. **c.** Total true positive molecules exist in screening data set across multiple benchmarking systems. Positive conformers represent molecules with different isomers. The summation of TPs are labeled across all benchmarked systems to calculate percentages in panel a.



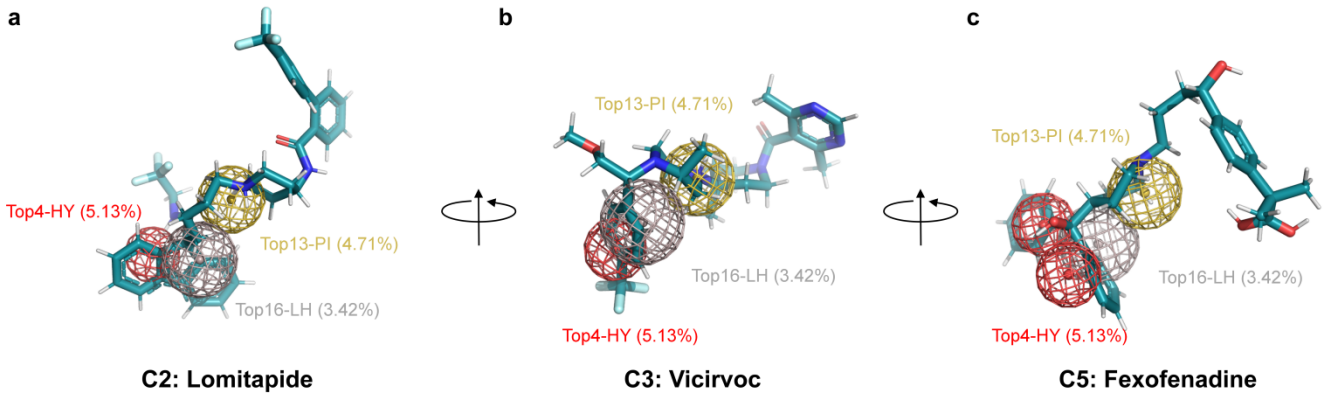
Supplementary Figure S12. Training and screening performance of Ph3DG on GPCR targets. **a, b, c.** Boxplots of AUROC, AUPRC and success rate (top 1%) across benchmarking methods, respectively, showing the comprehensive performance of Ph3DG in predicting ligand-protein bioactivity. Numbers in bracket indicate the number of targets successfully predicted by a specific method. Stars represent the outliers for each baselines. **d, e, f.** Statistics of recall rate, enrichment factor and ranking probability, respectively, showing the robust performance of Ph3DG in screening and retrieving positive compounds. In panel **d** and **e**, number in brackets indicate the number of GPCR targets successfully screened by a specific method, while for panel **f**, those numbers represent the number of TP compounds recalled by a specific method. Representative GPCRs tested in this study involves ADRB2, CCR5, CXCR4 and NK1R.



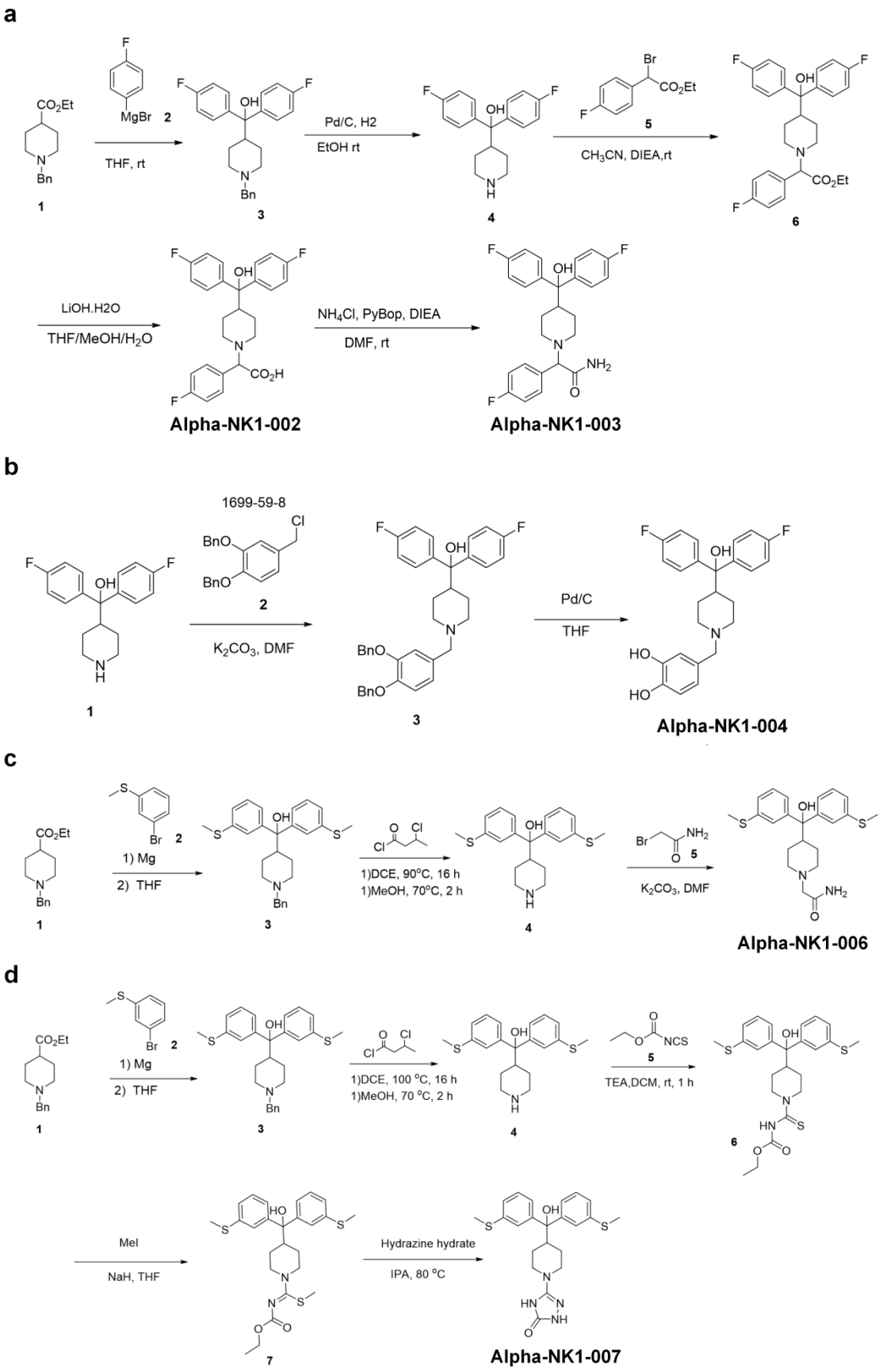
Supplementary Figure S13. Statistics of true-positives (TPs) used in benchmark studies. a. Total true positive molecules exist in screening data set across multiple benchmarking systems. Positive conformers represent molecules with different isomers.



Supplementary Figure S14. Screening protocol for NK1R and essential hit compounds with novel backbone. **a.** Screening processing is initialized with a comprehensive library comprising FDA-approved, clinical drug-like small molecule compounds. Ph3DG-MLP is applied to train pharmacophore model and screen compounds with more than 4 matching features. 12 compounds with computational and expertise inspection are tested *in vitro* resulting in 3 distinct bioactive hits. **b.** 2D chemical structures of bioactive hits (named Alpha-NK1-002, Alpha-NK1-003 and Alpha-NK1-005) and modified lead compounds (Alpha-NK1-005.1, Alpha-NK1-005.2, Alpha-NK1-005-f1 and Alpha-NK1-005-f5). Essential functional groups comprising pharmacophores are highlighted.



Supplementary Figure S15. PH4 patterns of NK1R screened compounds. Ligands are represented in turquoise stickers. Pharmacophore grid points are indiciated in mesh with color corresponding to **Fig. S8**. Fig. S16a: Alpha-NK1-002; b: Alpha-NK1-003; c: Alpha-NK1-005.



Supplementary Figure S16. Synthesis routes of NK1R optimized compounds. **a.** Synthesis of Alpha-NK1-002 and Alpha-NK1-003; **b.** Alpha-NK1-004; **c.** Alpha-NK1-006; **d.** Alpha-NK1-007.