

Supplementary Information for

**Chemical Genomics Language Model toward Reliable and
Explainable Compound-Protein Interaction Exploration**

*Takuto Koyama¹, Hayato Tsumura¹, Ryunosuke Okita², Kimihiro Yamazaki³, Aki Hasegawa¹,
Keiko Imamura^{4,5,6}, Takashi Kato³, Hiroaki Iwata⁷, Ryosuke Kojima^{1,8}, Haruhisa Inoue^{4,5,6},
Shigeyuki Matsumoto^{1*}, Yasushi Okuno^{1,9*}*

¹ Graduate School of Medicine, Kyoto University, 53 Shogoin-Kawaharacho, Sakyo-ku, Kyoto 606-8507, Japan

² Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto, 606-8501, Japan

³ Fujitsu Ltd., 4-1-1 Kamikodanaka, Nkahara-ku, Kawasaki, Kanagawa, 211-8588, Japan

⁴ Center for iPS Cell Research and Application (CiRA), Kyoto University, 53 Shogoin-Kawaharacho, Sakyo-ku, Kyoto 606-8507, Japan

⁵ Medical-Risk Avoidance Based On iPS Cells Team, RIKEN Center for Advanced Intelligence Project (AIP), Kyoto, Japan

⁶ iPSC-Based Drug Discovery and Development Team, RIKEN BioResource Research Center (BRC), Kyoto, Japan

⁷ Faculty of Medicine, Tottori University, 86 Nishi-cho, Yonago, Tottori 683-8503, Japan

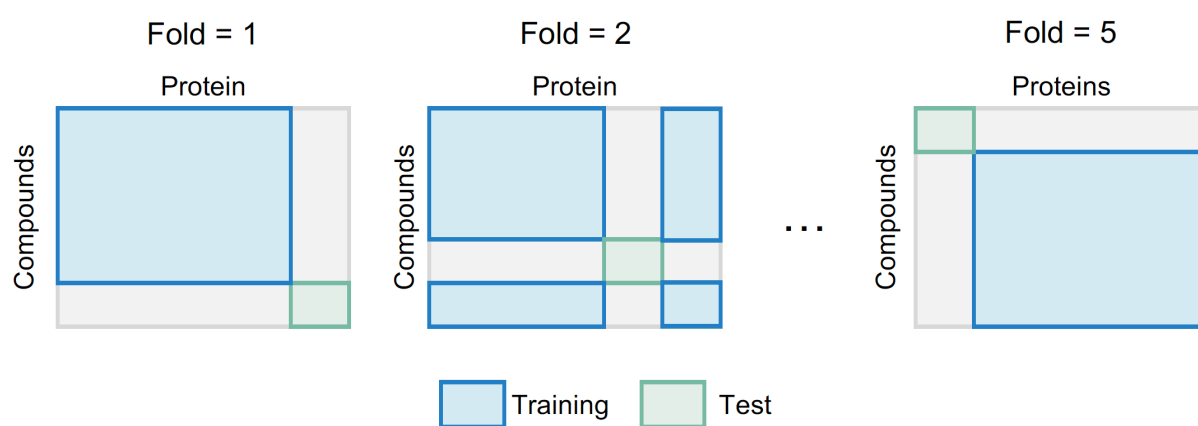
⁸ RIKEN Center for Biosystems Dynamics Research, 2-2-23 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

⁹ RIKEN Center for Computational Science, 7-1-26 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

* Corresponding authors

Supplementary Table S1: Summary of CPI datasets.

Name	Task	Compounds	Proteins	Interactions	Positives	Negatives
BindingDB	Classification	49,752	813	61,270	33,777	27,493
Davis	Classification	68	442	30,056	9,125	20,931
PDBbind	Regression	14,694	3,705	18,512	-	-
Metz	Regression	1,423	170	35,259	-	-



Supplementary Figure S1: Schematic illustration of zero-shot prediction. The whole dataset was divided into training and test sets so that both datasets don't share any compounds and proteins. The vertical axis of the matrices represents compounds, and the horizontal axis represents proteins.

Supplementary Table S2: Additional performance metrics for the BindingDB and Davis datasets.

Dataset	Method	Precision (Std)	Recall (Std)	BEDROC ($\alpha=20$)	BEDROC ($\alpha=80.5$)
BindingDB	NNS	-	-	0.611 (0.060)	0.623 (0.067)
	RandomForest	0.798 (0.242)	0.018 (0.014)	0.674 (0.066)	0.676 (0.067)
	XGBoost	0.800 (0.065)	0.497 (0.064)	0.673 (0.077)	0.674 (0.072)
	Komet	0.863 (0.027)	0.169 (0.068)	0.663 (0.036)	0.659 (0.039)
	GraphDTA	0.600 (0.121)	0.769 (0.214)	0.678 (0.108)	0.679 (0.123)
	TransformerCPI	0.723 (0.057)	0.499 (0.050)	-	-
	HyperAttentionDTI	0.800 (0.063)	0.579 (0.130)	0.727 (0.079)	0.728 (0.095)
	PerceiverCPI	0.748 (0.098)	0.423 (0.086)	0.726 (0.072)	0.737 (0.084)
	AttentionDTA	0.762 (0.074)	0.663 (0.015)	0.709 (0.098)	0.708 (0.116)
	Kang et al.	0.815 (0.068)	0.654 (0.049)	0.755 (0.072)	0.757 (0.105)
	ConPLex	0.801 (0.045)	0.675 (0.053)	0.694 (0.057)	0.697 (0.069)
	PSICHIC	0.768 (0.063)	0.580 (0.121)	0.686 (0.089)	0.671 (0.098)
	BarlowDTI	0.860 (0.012)	0.569 (0.062)	0.709 (0.091)	0.713 (0.097)
	ChemGLaM	0.808 (0.063)	0.714 (0.026)	0.722 (0.034)	0.709 (0.046)
Davis	NNS	-	-	0.417 (0.273)	0.418 (0.409)
	RandomForest	0.677 (0.092)	0.300 (0.127)	0.686 (0.110)	0.733 (0.143)
	XGBoost	0.559 (0.117)	0.357 (0.102)	0.533 (0.160)	0.531 (0.203)
	Komet	0.296 (0.048)	1.000 (0.000)	0.417 (0.031)	0.410 (0.033)
	GraphDTA	0.475 (0.129)	0.361 (0.212)	0.532 (0.147)	0.541 (0.169)
	TransformerCPI	0.511 (0.136)	0.417 (0.220)	-	-
	HyperAttentionDTI	0.405 (0.075)	0.638 (0.173)	0.560 (0.090)	0.590 (0.109)
	PerceiverCPI	0.510 (0.116)	0.393 (0.214)	0.578 (0.228)	0.597 (0.298)
	AttentionDTA	0.392 (0.045)	0.663 (0.173)	0.576 (0.206)	0.619 (0.251)
	Kang et al.	0.457 (0.040)	0.567 (0.163)	0.579 (0.105)	0.603 (0.128)
	ConPLex	0.382 (0.106)	0.258 (0.117)	0.403 (0.134)	0.394 (0.147)
	PSICHIC	0.463 (0.025)	0.475 (0.124)	0.485 (0.089)	0.503 (0.140)
	BarlowDTI	0.527 (0.124)	0.450 (0.243)	0.539 (0.138)	0.546 (0.140)
	ChemGLaM	0.561 (0.093)	0.466 (0.178)	0.625 (0.196)	0.654 (0.255)

BEDROC: Boltzmann-Enhanced Discrimination of ROC. α values indicated in column headers.

The reported BEDROC scores represent an 'average BEDROC'. This was calculated by first computing the BEDROC score for each target, and then averaging these target-specific scores to obtain the final score for each fold.

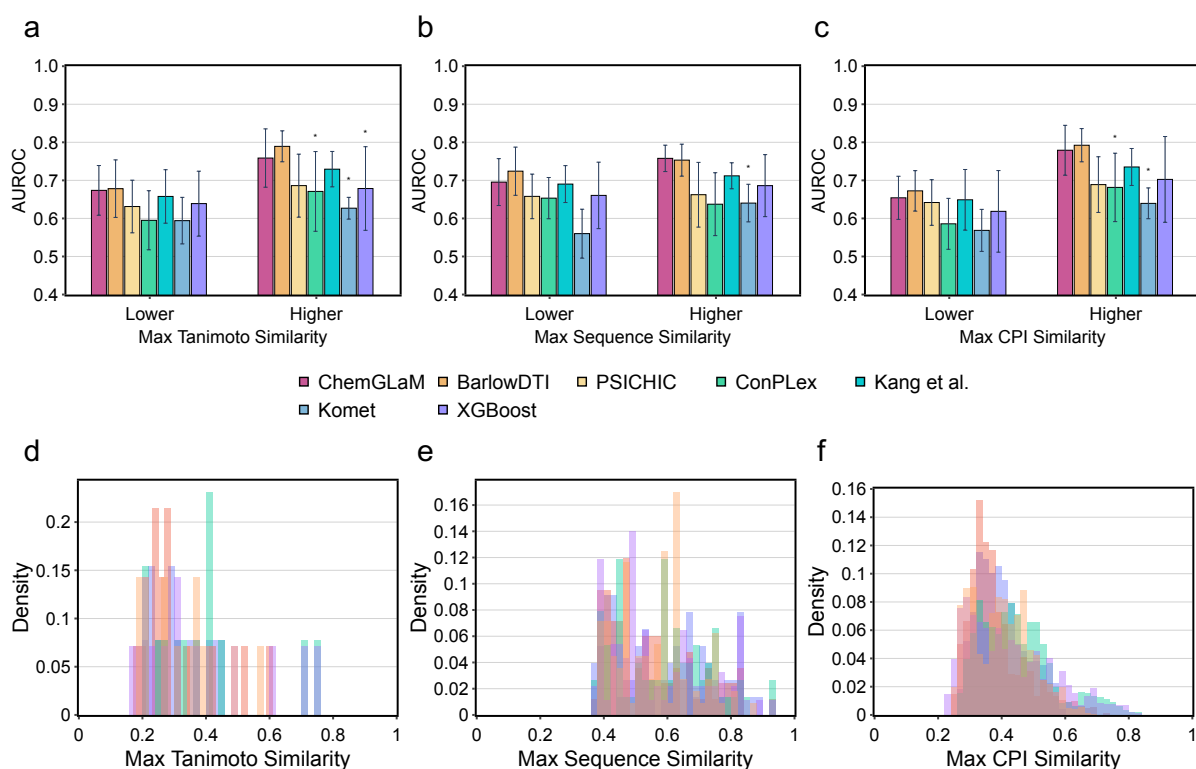
Std: Standard deviation from 5-fold cross-validation. Values in parentheses.

'-': BEDROC values omitted or not applicable for other models under the current BEDROC calculation method.

Bold: Best performance in the column for the dataset.

Underline: Second best performance in the column for the dataset.

Binary labels were determined by selecting the threshold that maximized the F1 score on the validation data.



Supplementary Figure S2: Performance evaluation of ChemGLaM in zero-shot prediction. **a-c)** The average AUROC scores of 5-fold cross-validation for the Davis dataset. The test set in each fold is divided into two groups based on the maximum similarities with the training set: close and distant pairs. The similarities are calculated via three approaches: Tanimoto, sequence, and CPI similarities. The CPI similarity is a harmonic mean of Tanimoto and sequence similarities. The bars on the left side represent the AUROC scores for the distant pairs (lower similarity) and those on the right side represent the scores for the close pairs (higher similarity). Statistical significance for comparisons was determined using a paired t-test with Benjamini/Hochberg FDR correction (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). **d-f)** Distributions of maximum Tanimoto, sequence, and CPI similarities in five folds are represented. The distributions in each fold are indicated by different colors.

Supplementary Table S3: Performance Evaluation on BioSNAP Cold Splits

Scenario	Method	AUROC (Std)	AUPRC (Std)
Unseen Drug	DeepDTI	0.843 ± 0.003	-
	DeepDTA	0.849 ± 0.007	-
	DeepConv-DTI	0.847 ± 0.009	-
	MolTrans	0.853 ± 0.011	-
	PSICHIC	0.8833	0.9018
	BarlowDTI	0.9563	0.9630
	ChemGLaM	0.8469	0.8650
	ChemGLaM-opt	0.8748	0.8922
Unseen Protein	DeepDTI	0.759 ± 0.029	-
	DeepDTA	0.767 ± 0.022	-
	DeepConv-DTI	0.766 ± 0.022	-
	MolTrans	0.770 ± 0.029	-
	PSICHIC	0.8681	0.8685
	BarlowDTI	0.9466	0.9493
	ChemGLaM	0.7730	0.7864
	ChemGLaM-opt	0.7888	0.7970

AUROC: Area under the receiver operating characteristic curve.

AUPRC: Area under the precision-recall curve.

Std: Standard deviation reported by original methods where available. - indicates not reported or not applicable.

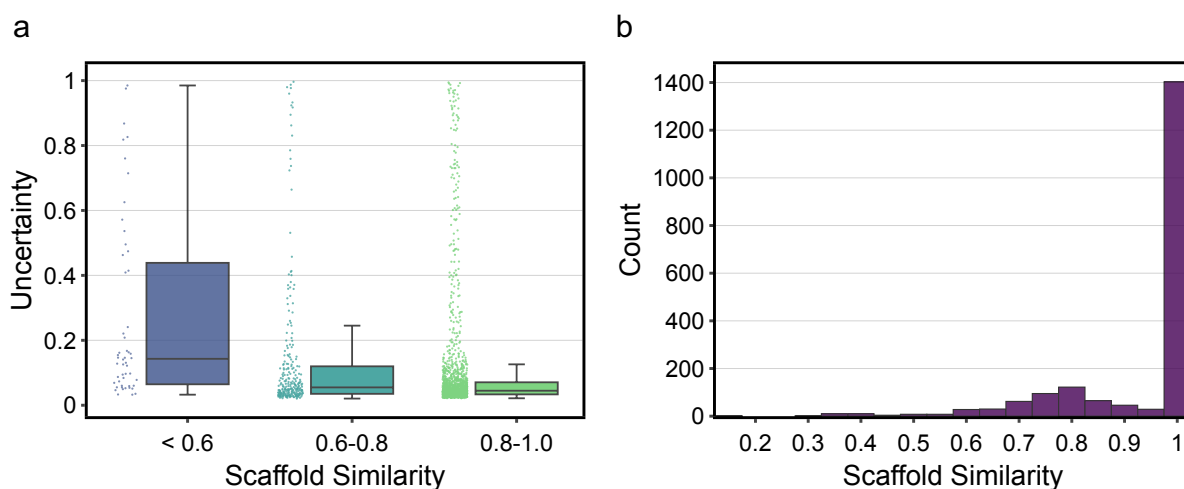
ChemGLaM-opt: ChemGLaM with optimized hyperparameters.

Supplementary Table S4: Hyperparameters Explored for ChemGLaM-opt

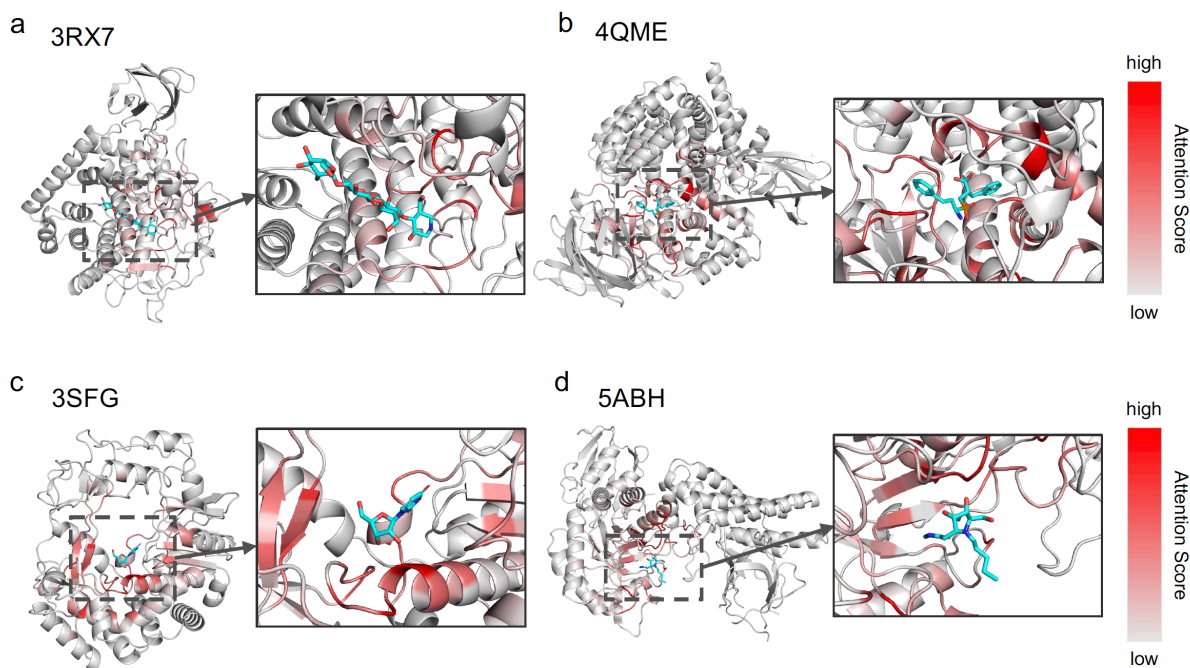
Hyperparameter	Values Explored
Learning Rate	1×10^{-5} , 5×10^{-6} , 1×10^{-6}
Batch Size	8, 16, 32
Epochs	10, 20
Dropout Rate	0.1, 0.2

S1 Additional analysis of uncertainty estimation using EDL

To specifically assess the relationship between uncertainty and chemical novelty (i.e., scaffold hops), we performed an evaluation using EDL on the 'unseen compound' data split (where proteins are shared but compounds are unseen, corresponding to the gray area in Supplementary Figure S1). Scaffold Similarity was defined on a per-target basis by calculating the maximum Bemis-Murcko scaffold similarity against the training data.



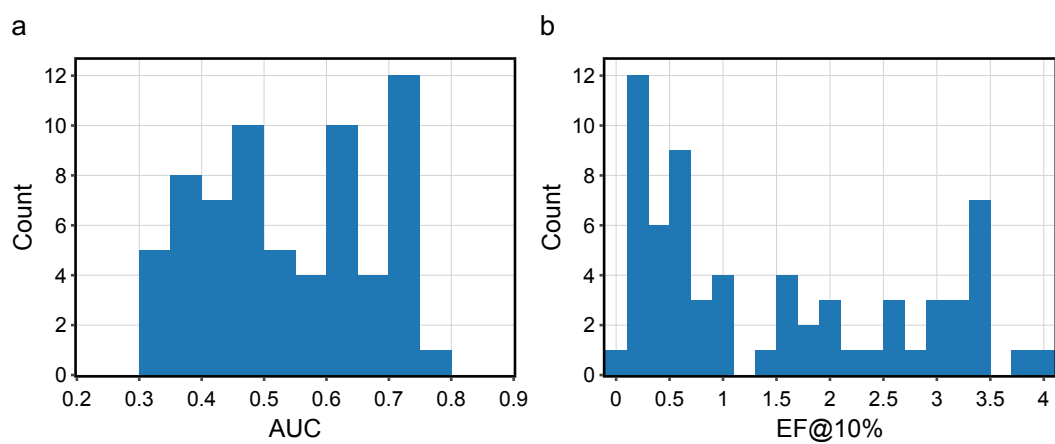
Supplementary Figure S3: Analysis of chemical novelty versus prediction uncertainty on the BindingDB 'unseen compound' split. **a)** Distributions of uncertainty scores for test set compounds, grouped by their maximum Bemis-Murcko scaffold similarity to the training set. **b)** Histogram showing the distribution of the scaffold similarities.



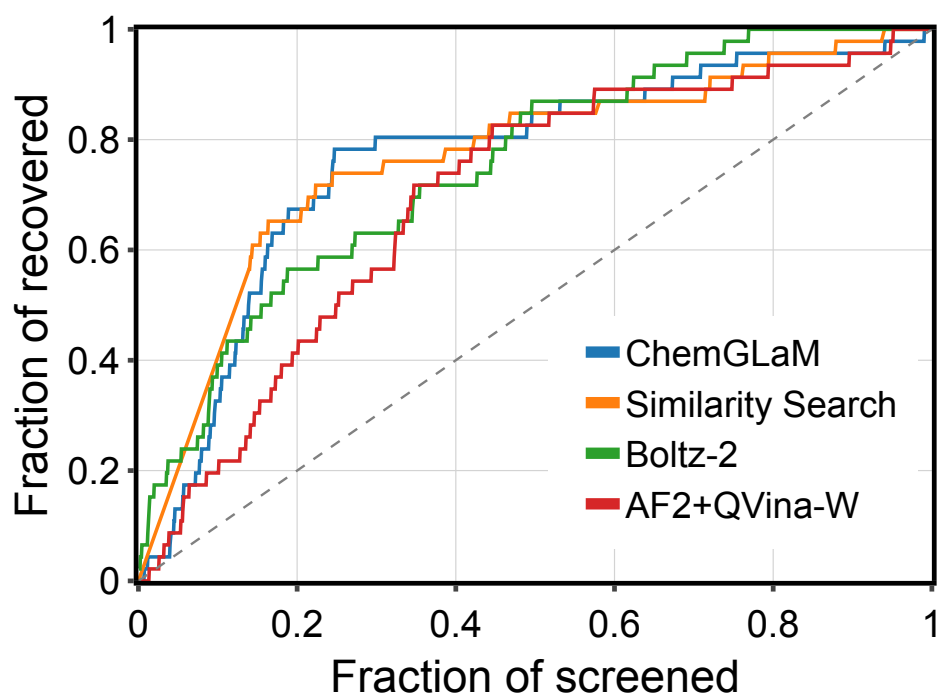
Supplementary Figure S4: Visualizing attention maps on other complex structures. Attention visualization for **a)** AaCel9A in complex with cellotetraose-like isofagomine (PDB entry = 3RX7), **b)** Aminopeptidase N in complex with the phosphinic dipeptide analogue LL-(R, S)-hPheP[CH₂]Phe (PDB entry = 4QME), **c)** murine norovirus RNA dependent RNA polymerase in complex with 2thiouridine(2TU) (3SFG), and **d)** GH84 with ligand (PDB entry = 5ABH). Max-pooled attention scores, which were min-max normalized along the sequence direction, were assigned as B-factors. A color scale ranging from gray90 to red was then applied based on these B-factors. The color scale was set with a minimum value of 0.2 and a maximum value of 1.0.

Supplementary Table S5: Protein list used in current study as related to Src/c-Abl pathway

UniProt ID	Protein names
P00519	Tyrosine-protein kinase ABL1
Q9UM73	ALK tyrosine kinase receptor
P11362	Fibroblast growth factor receptor 1
P21802	Fibroblast growth factor receptor 2
P22607	Fibroblast growth factor receptor 3
P22455	Fibroblast growth factor receptor 4
P17252	Protein kinase C alpha type
P05771	Protein kinase C beta type
P05129	Protein kinase C gamma type
P24723	Protein kinase C eta type
P08581	Hepatocyte growth factor receptor
P16234	Platelet-derived growth factor receptor alpha
P09619	Platelet-derived growth factor receptor beta
P12931	Proto-oncogene tyrosine-protein kinase Src
P17948	Vascular endothelial growth factor receptor 1
P35968	Vascular endothelial growth factor receptor 2
P35916	Vascular endothelial growth factor receptor 3



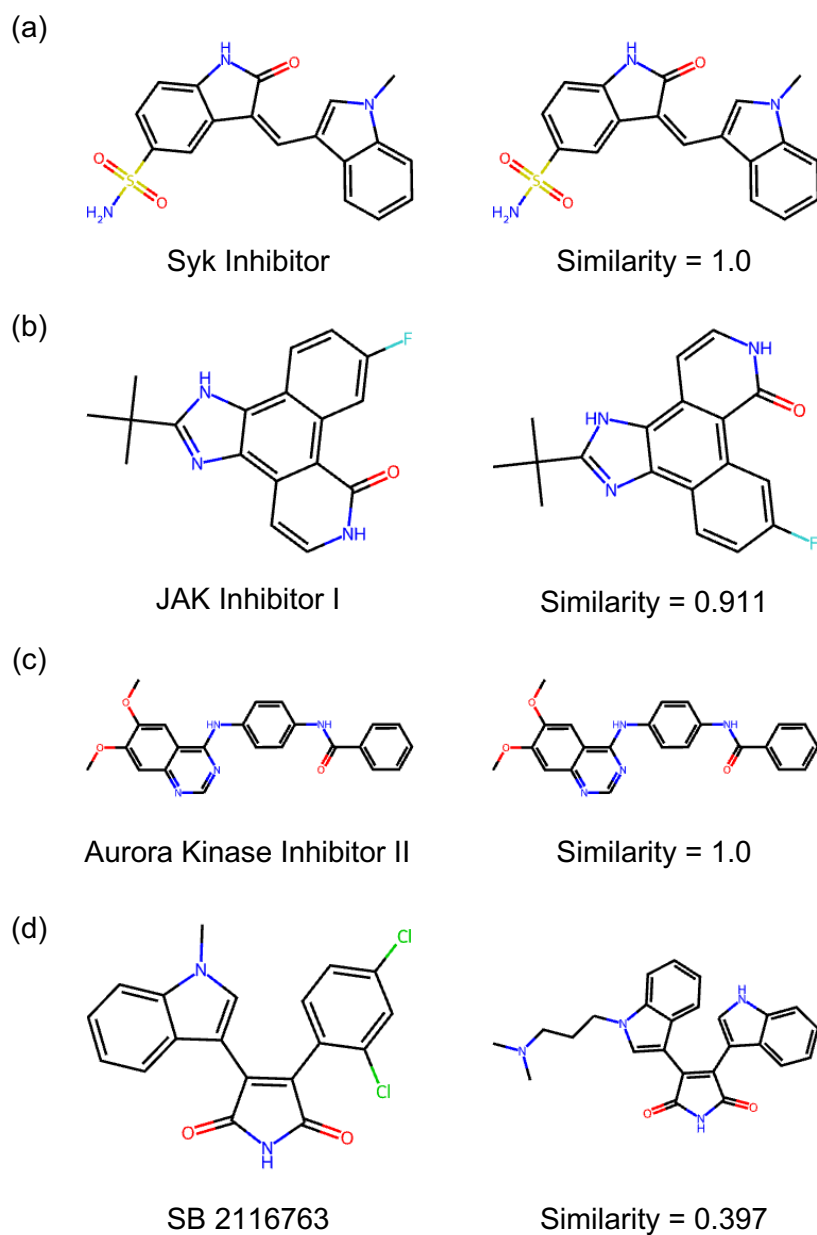
Supplementary Figure S5: Distribution of **a)** AUC values and **b)** enrichment factors for hit enrichment curves of 66 reported targets in ALS treatment ²⁹.



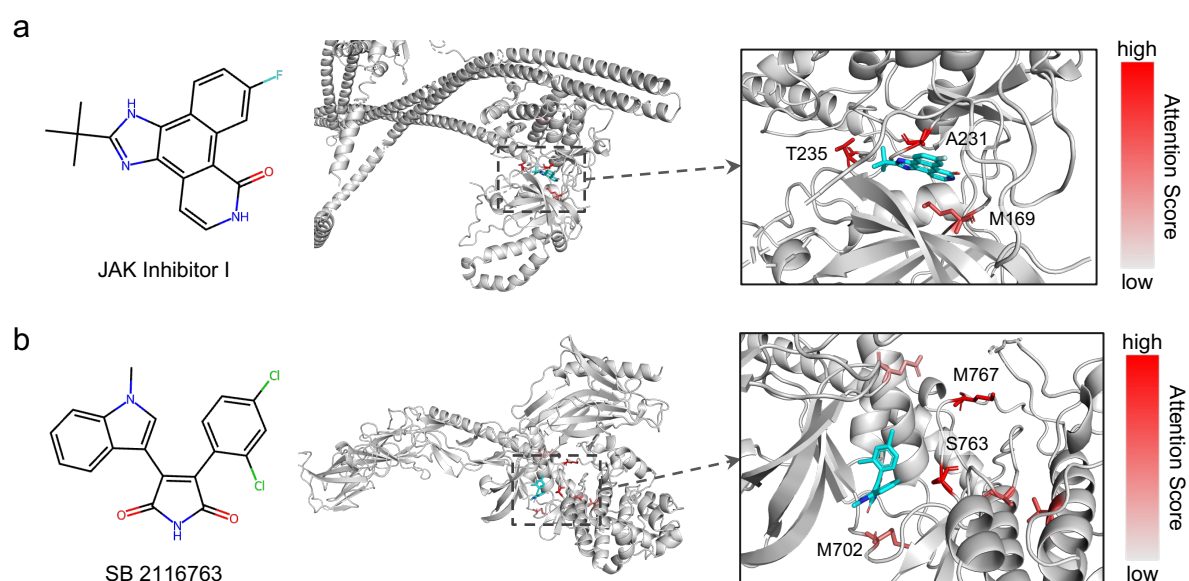
Supplementary Figure S6: Enrichment analysis of Src/c-abl pathways using drug screening data on iPS cells from patients with SOD1-ALS. The enrichment curve was plotted against the maximum scores of cross-protein output scores.

Supplementary Table S6: Performance metrics for enrichment analysis.

Method	AUC	EF@1%	EF@5%	EF@10%	EF@20%
Similarity Search	0.767	4.12	3.93	3.53	2.96
Vina	0.694	0.00	1.75	1.97	2.07
Boltz-2	0.751	6.55	4.37	3.93	2.83
ChemGLaM	<u>0.763</u>	2.18	2.62	3.28	3.37



Supplementary Figure S7: a-d) Nearest neighbor analysis of hit compounds with the highest MNIs in the additional data.



Supplementary Figure S8: a) Attention visualization for JAKInhibitor I and its predicted target, ROCK2.
e) Attention visualization for SB 2116763 and its predicted target, EPHA4. The complex structures were predicted by HelixFold3. The color scale was set with a minimum value of 0.2 and a maximum value of 1.0.