SUPPLEMENTARY MATERIAL
BiGMINT: Biologically-guided Hierarchical Multimodal Integration for Modeling Multiple Compound Activities in Drug Discovery

Pushpak Pati† Bo Li† Abbas Rayabat Khan Tomé Albuquerque Steffen Jaensch
Amina Mollaysa Walid M. Abdelmoula Samantha J. Allen Joke Reumers Helai P. Mohammad
Scott Oloff Tommaso Mansi Rui Liao Dmytro S. Lituiev Zhoubing Xu
Janssen R&D, LLC, a Johnson & Johnson Company
ppati4@its.jnj.com

S.1. Notations

Data notations	
$\mathcal{D}$	compound activity dataset
$\mathcal{C}$	set of compounds
$\mathcal{P}$	set of proteins
c	a single compound $\in \mathcal{C}$
p	a single protein $\in \mathcal{P}$
$n \in N_c$	number of HCI images for compound c
$\mathbf{x}_{nc} \in \mathbb{R}^{h \times w \times ch}$	single image patch for compound c
$\{\mathbf{x}_{nc}\}_{n=1}^{N_c}$	set of image patches for compound c
Pharmacological setup notations	
A_p	a pharmacological assay performed to test compound-induced modulation targeting protein p
$\mathcal{Z}_p$	set of concentrations at which assay A_p is measured
z	a single assay-specific concentration $\in \mathcal{Z}_p$
$\mathcal{T}_p$	set of tasks denoting measurements for assay A_p across concentrations $\mathcal{Z}_p$
$t_{p,z}$	a single task $\in \mathcal{T}_p$ measuring A_p at concentration $z \in \mathcal{Z}_p$
Ground-truth binary activity notations	
$\mathcal{Y}_c$	set of ground-truth binary activities for compound c measured across all proteins $p \in \mathcal{P}$ and assay-specific concentrations $z \in \mathcal{Z}_p$
$y_{c,(p,z)}$	single binary activity of compound c measured for protein p at concentration z

Table 1. List of notations and their descriptions used in the paper.

† These authors contributed equally to this paper.

S.2. Datasets

Compound activity datasets: See Table 2 for the exact number of compounds used for evaluating the compound activity modeling performance. The U2OS and iNeuron datasets included ∼99K and ∼40K unique compounds, respectively, over an order of magnitude more compounds for activity modeling than recent benchmarks [12, 13, 17], enabling robust analysis. The inclusion of multiple experimental batches in these datasets also facilitated a thorough assessment of batch variation effects in HCI [2].

The datasets were acquired using Cell Painting on U2OS and iNeuron cell lines with 24-hour incubation and compound concentrations of 10μM and 20μM , respectively. Different cell components were labeled using fluorescent dyes and acquired 16-bit 5-channel fluorescence images with a Yokogawa CellVoyager 8000 confocal HCI reader at 20× magnification [15]. In total, 170 binary activity tasks were measured across 65 proteins and 123 protein-based assays, with certain proteins assayed using multiple protocols. The compound-to-task matrix was sparsely populated with fill rates of **2.94%** for U2OS and **3.01%** for iNeuron, corresponding to ∼522K and ∼475K activity annotations, respectively. Per-task fill rates were $2.94 \pm 8.45\%$ for U2OS and $3.01 \pm 8.77\%$ for iNeuron. For evaluation, datasets were partitioned into 5-folds at scaffold-level using Tanimoto similarity [15].

Figure 1 presents box plots illustrating the ratio of positive to negative activity labels across the 5-folds for the U2OS and iNeuron datasets. The medians of the box plots (indicated by red markers) reveal substantial class imbalance in most tasks. Furthermore, the wide ranges of the quantiles highlight considerable variation in label distribution across folds. These imbalanced class distributions justify the decision of stacking the predictions across all test folds before computing quantitative metrics, rather than cal-

Dataset	Usage	#compounds	#batches	#plates	#wells	#FoVs	FoV shape
U2OS	FM pretrain	119,876	87	1,364	353,607	353,607	$970 \times 970 \times 5$
iNeuron	FM pretrain	73,101	48	308	95,908	862,757	$1938 \times 1938 \times 5$
U2OS	BiGMINT evaluation	99,071	318	560	186,776	746,970	$970 \times 970 \times 5$
iNeuron	BiGMINT evaluation	39,444	48	627	53,757	483,607	$1938 \times 1938 \times 5$

Table 2. U2OS and iNeuron HCI dataset statistics for pre-training FMs and evaluating the BiGMINT compound activity framework.

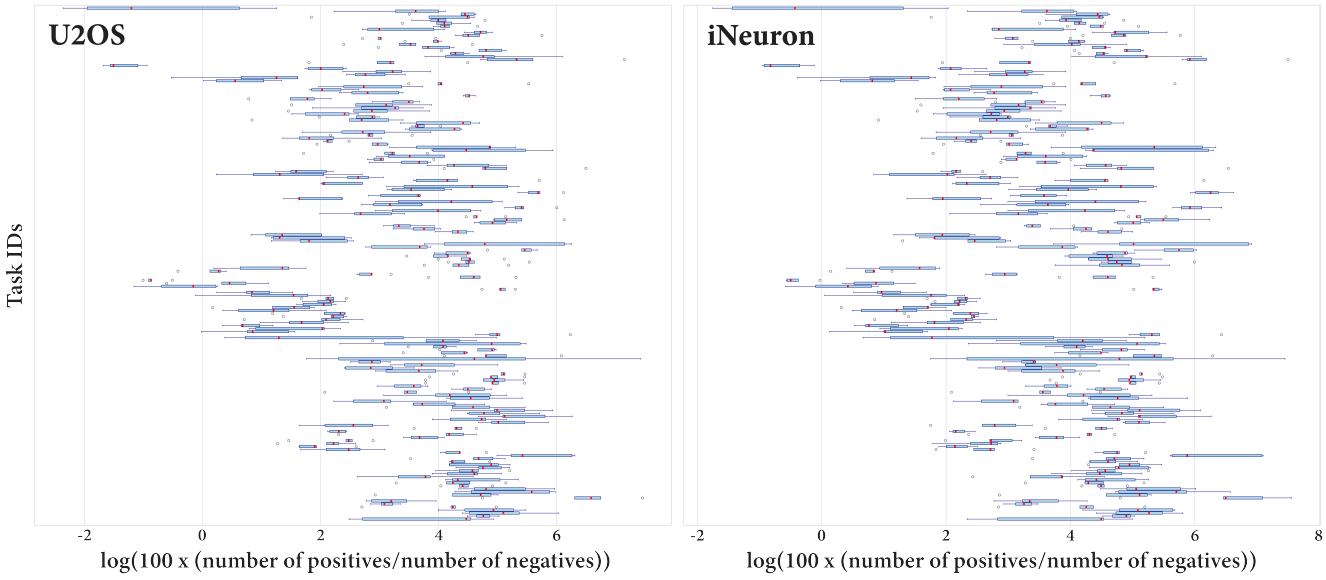

Figure 1. Per-task ratio of positive to negative activity labels across the 5-folds in U2OS and iNeuron compound activity datasets. Note that the positive-to-negative ratio can vary substantially between folds and between tasks.

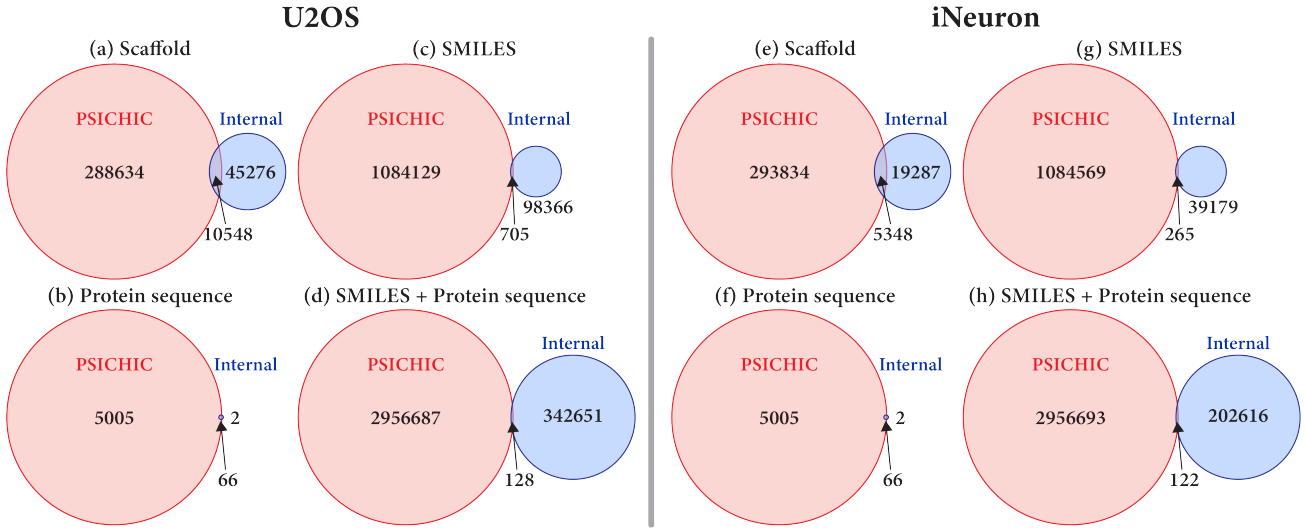

Figure 2. Compound, protein, and compound–protein pair overlaps between PSICHIC pretraining and our internal activity datasets.

culating metrics per fold and reporting aggregated statistics.

External evaluation: At present, no public benchmark jointly provides the three modalities required for target-directed compound activity modeling in our setting: protein targets, compound structures, and HCI. Large phenotypic datasets such as JUMP-CP [7], RxRx1/3 [20], and cpg0036 [35] lack explicit protein target annotations, while bioactivity repositories including ChEMBL [37], BindingDB [23], PDB [24], and Boltz2 [27] do not include matched HCI readouts. Furthermore, identifying protein

targets from assay descriptions as in [25] is insufficient, as most assays are functional or pathway-level rather than direct drug–target interaction measurements. Consequently, comprehensive public tri-modal benchmarks suitable for external evaluation are currently unavailable. By systematically demonstrating the complementary value of protein targets, compound structures, and HCI for drug–target interaction modeling, this work motivates the development of future public tri-modal benchmarks and provides a scalable methodological foundation for advancing multimodal learning in this domain.

Cell line–agnostic Chemoproteomic Pretraining: The encoder $\mathcal{F}_\alpha$ in $\mathcal{F}^{\text{chemprot}}$ was initialized with pretrained PSICHIC [19] weights. PSICHIC was trained using a large, diverse, experimentally validated dataset to jointly predict 2,956,815 molecular binding affinities (pKi) and 1,842,394 functional effects (agonists, antagonists, non-binders) between 5,071 unique proteins and 1,084,834 unique compounds. As pKi reflects binding affinity and is not directly equivalent to compound activity, transfer learning from PSICHIC is justified without information leakage. For transparency, we report compound, protein, and compound–protein pair overlaps between PSICHIC pretraining and our activity datasets in Figure 2.

While 66 out of 68 proteins in our dataset from U2OS and iNeuron cell lines also appear in the pretraining dataset from PSICHIC, such overlap is expected given that both focus on well-studied, pharmacologically important, and clinically relevant drug targets prevalent in public bioassay resources [5, 30]. This high protein overlap reflects the shared biological focus of drug discovery rather than direct information leakage. Compound overlap between our dataset and the PSICHIC pretraining dataset is low (0.7%), and identical protein–compound pairs are rare ($0.03\% - 0.06\%$), suggesting that our compound activity prediction task against known targets is unlikely to be significantly influenced by pretraining exposure.

Cell line-specific HCI Pretraining: Table 2 displays the detailed number of compounds, batches, plates, wells, and FoVs for pretrain the cell line-specific HCI models. The U2OS HCI FM $\mathcal{F}_\omega$ was trained on the public cell painting dataset from the JUMP-CP consortium [7], specifically cpg0016-jump dataset [34]. The pretraining subset comprised $\sim$119K diverse compounds at 10μM, imaged at 20$\times$ magnification, yielding $\sim$2.8M 16-bit 5-channel FoVs. To accelerate FM training while maintaining diversity, we randomly selected one FoV per well, resulting in $\sim$353K FoVs.

The iNeuron HCI FM $\mathcal{F}_\omega$ was trained on an internal dataset of $\sim$73K compounds on iNeuron cell lines at 20μM. Imaging with a Yokogawa CellVoyager 8000 confocal HCI reader at 20$\times$ magnification [15] produced $\sim$862K 16-bit, 5-channel fluorescence FoVs.

S.3. Method

S.3.1. Chemoproteomic Interaction Encoder

$\mathcal{F}^{\text{chemprot}}$ consists of two components $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$. $\mathcal{F}_\alpha$ takes the protein sequence $\mathbf{p}$ and compound SMILES $\mathbf{c}$ as input, and outputs the molecular interaction signal between the protein and compound. This signal is then projected to $\mathbf{d}_{c,p}^{\text{chemprot}} \in \mathbb{R}^d$ by $\mathcal{F}_\beta$. In our paper, $\mathcal{F}_\alpha$ is initialized with the pretrained PhySIcoCHemICal Graph Neural Network (PSICHIC) and $\mathcal{F}_\beta$ is a MLP, which we explain below.

S.3.1.1. PSICHIC

PSICHIC [19] incorporates physicochemical constraints to decode interaction fingerprints directly from sequence data alone. More specifically, an input compound $\mathbf{c}$ and protein $\mathbf{p}$ are converted into 2D molecular and protein graphs, $(\mathrm{H}_c^0, \mathrm{E}_c^0, \mathrm{A}_c)$ and $(\mathrm{H}_p^0, \mathrm{E}_p^0, \mathrm{A}_p)$, via RdKit [4], ESM2 [22] and physiochemical properties, where $\mathrm{H}, \mathrm{E}, \mathrm{A}$ denote node features, edge features, and adjacency matrices.

These graphs are processed through a stack of Physicochemical Graph Convolution (PGC) layers. Each layer l sequentially, (i) models **intra-molecular forces** (Eq. 1) to propagate information among neighboring atoms or residues; (ii) applies **physicochemical constraints** (Eq. 2) to group atoms into functional groups and residues into protein regions; and (iii) models **inter-molecular forces** (Eq. 3) to promote cross-interaction between compound and protein regions before updating node features to layer $l+1$. See [19] for the details of $\mathrm{INTRA_C}, \mathrm{INTRA_P}, \mathrm{CONSTR_C}, \mathrm{CONSTR_P}$, and INTER operations.

$$\begin{aligned} \mathrm{H}_{c,\text{intra}}^{l+1} &= \mathrm{INTRA}_\mathrm{C}^l(\mathrm{H}_c^l, \mathrm{E}_c^l, \mathrm{A}_c) \\ \mathrm{H}_{p,\text{intra}}^{l+1} &= \mathrm{INTRA}_\mathrm{P}^l(\mathrm{H}_p^l, \mathrm{E}_p^l, \mathrm{A}_p) \end{aligned} \tag{1}$$

$$\begin{aligned} \mathrm{H}_{c,\text{fgroup}}^{l+1}, \mathrm{G}_{c,\text{fgroup}}^{l+1} &= \mathrm{CONSTR}_\mathrm{C}^l(\mathrm{H}_{c,\text{intra}}^{l+1}, \mathrm{A}_c) \\ \mathrm{H}_{p,\text{region}}^{l+1}, \mathrm{G}_{p,\text{region}}^{l+1} &= \mathrm{CONSTR}_\mathrm{P}^l(\mathrm{H}_{p,\text{intra}}^{l+1}, \mathrm{A}_p) \end{aligned} \tag{2}$$

$$\begin{aligned} \mathrm{H}_c^{l+1}, \mathrm{H}_p^{l+1}, = \mathrm{INTER}^l(&\mathrm{H}_{c,\text{intra}}^{l+1}, \mathrm{H}_{c,\text{fgroup}}^{l+1}, \mathrm{G}_{c,\text{fgroup}}^{l+1}, \\ \mathrm{M}_c^{l+1}, \mathrm{M}_p^{l+1} \quad &\mathrm{H}_{p,\text{intra}}^{l+1}, \mathrm{H}_{p,\text{region}}^{l+1}, \mathrm{G}_{p,\text{region}}^{l+1}) \end{aligned} \tag{3}$$

M_c^{l+1} and M_p^{l+1} capture the significance of atoms and protein residues in the formation of compound-protein interactions. The final interaction fingerprint $\mathbf{f}$ after L PGC layers is obtained from Eq. 4:

$$\begin{aligned} \mathrm{M}_c^{\text{pool}} &= \mathrm{softmax}([\mathrm{M}_c^1 \,||\, \ldots \,||\, \mathrm{M}_c^L].\mathrm{W_C}) \\ \mathrm{M}_p^{\text{pool}} &= \mathrm{softmax}([\mathrm{M}_p^1 \,||\, \ldots \,||\, \mathrm{M}_p^L].\mathrm{W_P}) \\ \mathbf{f} &= \left[\mathrm{MLP_P}(\mathrm{M}_p^{\text{pool}}.\mathrm{H}_p^L) \,||\, \mathrm{MLP_C}(\mathrm{M}_c^{\text{pool}}.\mathrm{H}_c^L)\right] \\ \mathbf{q} &= \mathrm{M}_p^{\text{pool}} \cdot \mathrm{H}_p^L || \mathrm{M}_c^{\text{pool}} \cdot \mathrm{H}_c^L \end{aligned} \tag{4}$$

Method	U2OS			iNeuron		
	AUCROC	AUPRC	Macro F1	AUCROC	AUPRC	Macro F1
$\mathbf{f} \rightarrow \text{pIC}_{50}$	67.40	44.70	54.00	70.90	52.49	55.90
$\mathbf{q} \rightarrow \text{pIC}_{50}$	<u>71.11</u>	<u>48.29</u>	<u>58.75</u>	<u>72.62</u>	<u>54.22</u>	<u>57.83</u>

Table 3. Compound activity prediction performance (%) for using different chemoproteomics-embedding for binding strength prediction. Using the intermediate embedding $\mathbf{q}$ instead of PSICHIC's fingerprint $\mathbf{f}$ results in better quantitative performance.

where $||$ denotes concatenation. $\text{M}_c^{\text{pool}} \in \mathbb{R}^{1\times \text{N}_c}$ and $\text{M}_p^{\text{pool}} \in \mathbb{R}^{1\times \text{N}_p}$, where N_c is the number of atoms and N_p is the number of protein residues. M_c^{pool} and M_p^{pool} capture the aggregated significance of atoms and residues, respectively. They are computed by concatenating significance scores across L PGC layers, followed by MLP projections. The protein and compound embeddings from the L^{th} PGC layer, H_p^L and H_c^L, are combined with the aggregated significance vectors, M_p^{pool} and M_c^{pool}, followed by MLPs and concatenation to generate the interaction fingerprint $\mathbf{f}$.

For BiGMINT, we consider an alternative embedding, $\mathbf{q}$, as the molecular interaction signal. Specifically, we combine the weighted protein and compound embeddings from the L^{th} PGC layer, $\text{M}_p^{\text{pool}} \cdot \text{H}_p^L$ and $\text{M}_c^{\text{pool}} \cdot \text{H}_c^L$, and concatenate them to yield $\mathbf{q}$. We prefer $\mathbf{q}$ over $\mathbf{f}$ because $\mathbf{f}$ is explicitly optimized for pKi prediction and is therefore highly specific to that endpoint. In contrast, our task of predicting pIC_{50} relies on related but distinct molecular relationships. Consequently, the intermediate features $\mathbf{q}$ can provide more generalizable and transferable molecular interaction signals. Empirical differences in performance between $\mathbf{q}$ and $\mathbf{f}$ for compound activity prediction are presented in Table 3. Finally, the molecular interaction signal $\mathbf{q}$ is projected via $\mathcal{F}_\beta$, such that $\mathcal{F}_\beta(\mathbf{q}) \in \mathbb{R}^d$, to guide HCI-feature extraction.

S.3.2. Chemoproteomic-Guided HCI Encoder

$\mathcal{F}^{\text{hci}}$ consists of three components, $\mathcal{F}_\omega$, $\mathcal{F}_\phi$, and $\mathcal{F}_\Psi$, which encodes HCI into phenotypic embedding $\mathbf{d}^{\text{hci}} \in \mathbb{R}^{|\mathcal{T}|\times d}$. $\mathbf{d}^{\text{hci}}$ comprises embeddings $\text{d}^{\text{hci}}_{c,(p,z)}$ related to tasks $t_{c,(p,z)}$.

$\mathcal{F}_\omega$ is a foundation model (FM) pretrained with self-supervision on HCI from a disjoint compound dataset $\mathcal{D}^*$, *s.t.* $\mathcal{D} \cap \mathcal{D}^* = \emptyset$. Specifically, U2OS and iNeuron FoVs are resized to 960×960, followed by channel-wise intensity clipping at the < 0.01 and > 99.9 percentiles, and min-max normalization to improve signal-to-noise ratio. Then, ViT-B/16 [11] is trained with DINOv2 [6] for U2OS and DINO [26] for iNeuron. Empirically, we observed inferior representation quality with DINOv2 for iNeuron, likely due to the inadequacy in reconstructing fine neurite structures. The FMs are trained for 400 epochs using two 480×480 global crops and eight 128×128 local crops. For effective knowledge distillation, crops are drawn from high-intensity nuclear regions and augmented with flips, rotations, blur and color jitter. During inference, FoVs are partitioned into 480×480 image patches, embedded into 768-dimensions by FMs, and plate-wise batch corrected using robust z-scoring against control image patch statistics.

$\mathcal{F}_\phi$ is a shared MLP projector. $\mathcal{F}_\Psi$ is a collection of phenotypic aggregators $\{\mathcal{F}_{\psi_{p,z}} \mid p \in \mathcal{P},\ z \in \mathcal{Z}_p\}$. Each head $\mathcal{F}_{\psi_{p,z}}$ employs cross-attention between chemoproteomic embedding $\text{d}^{\text{chemprot}}_{c,p}$ as query and projected patch features from $\mathcal{F}_\phi$ as keys and values, producing image embedding $\text{d}^{\text{hci}}_{c,(p,z)}$ for task $t_{p,z}$. Note that we also investigated sharing the cross-attention module across phenotypic aggregators to improve parameter efficiency. However, this parameter sharing diminished the effectiveness of per-task adaptation, in terms of chemoproteomics-guidance and HCI feature aggregation, leading to empirically observed reductions in activity modeling performance across tasks. Shared and not-shared cross-attention resulted in mean AUCROC 77.1% vs 75.8% on U2OS dataset, and 74.9% vs 73.8% on iNeuron dataset.

S.3.3. Cross-modal Fusion

Since different protein mechanisms and measured tasks induce different cross-modal structures, we evaluated multiple cross-modal fusion operators within $\mathcal{F}^{\text{fusion}}_{p,z}$ to find the best bias for molecular–phenotypic coupling, including concatenation [9, 33], vector-wise gating [38], element-wise gating [1], attention [14], outer-product [8], to probe different interaction inductive biases. For instance, concatenation captures additive complementarity; gating enables conditional modulation of one modality by the other; attention supports sparse, selective reuse of modality information; and the outer-product encodes multiplicative, pairwise couplings between embedding dimensions.

Let $\text{d}^{\text{chemprot}}_{c,p}$ and $\text{d}^{\text{hci}}_{c,(p,z)}$ are chemoproteomics and HCI embeddings. W and W_g are learnable linear projections, $||$ denotes vector concatenation, $\sigma(\cdot)$ is sigmoid nonlinearity, $\odot$ is element-wise multiplication, and $\text{CA}(Q, K)$ returns cross-attention signal from query Q over keys K. Formally the cross-modal fusion techniques are presented in Eqn. 5.

More specifically, for $f_{\text{gating-ChemProt}}$, $\text{d}^{\text{hci}}_{c,(p,z)}$ serves as the base representation, while $\text{d}^{\text{chemprot}}_{c,p}$ is used as the conditioned modality. Conversely, for $f_{\text{gating-HCI}}$, these roles are reversed; that is, $\text{d}^{\text{chemprot}}_{c,p}$ acts as the base and $\text{d}^{\text{hci}}_{c,(p,z)}$

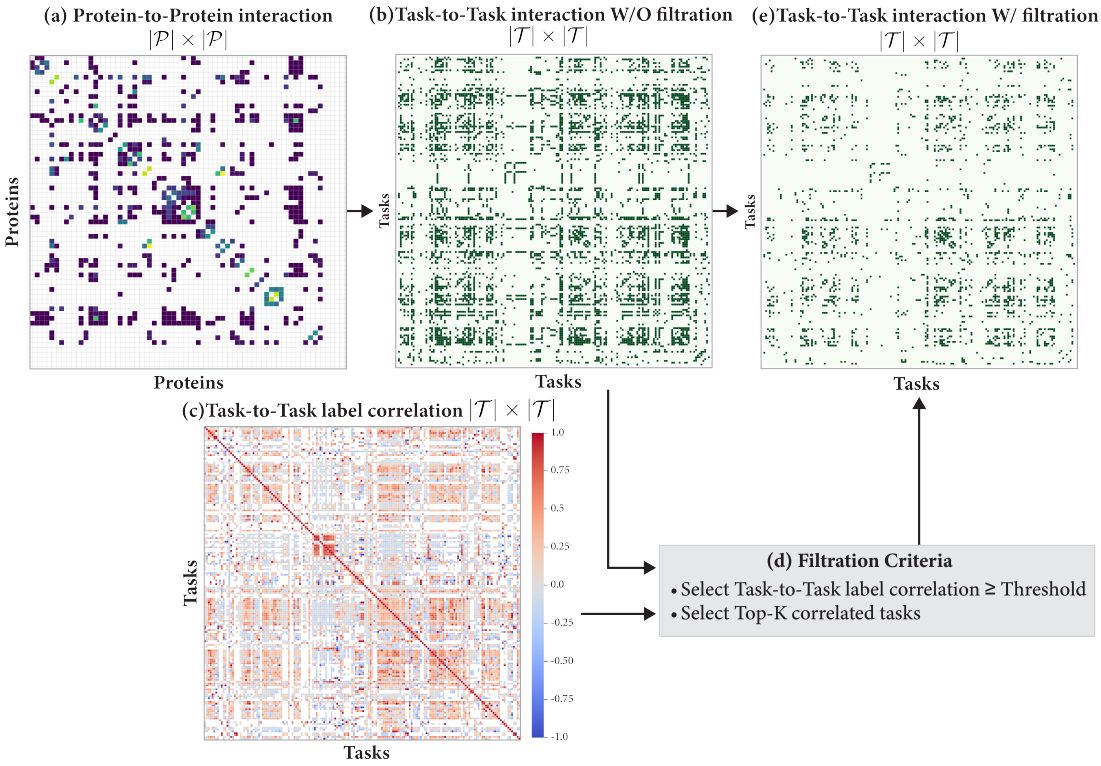

Figure 3. Overview of the workflow for deriving biological prior PPI and TTI for facilitating PPI-guided task embedding augmentation.

is the conditioned modality. In Figure 4(c) of the main paper, "Gated Add" and "Gated Add (reverse)" correspond to $f_{\text{gating-ChemProt}}$ and $f_{\text{gating-HCI}}$, respectively.

Element-wise gating is implemented by modifying the gate definition. W_g produces $\text{gate} \in \mathbb{R}^d$, enabling element-wise gating of chemoproteomics and HCI features within $f_{\text{gating-ChemProt}}$ and $f_{\text{gating-HCI}}$. The operations "Gated element Add" and "Gated element Add (reverse)" are defined analogously to their non-element-wise counterparts.

$$
\begin{aligned}
f_{\text{concat}} &= W\left[\mathrm{d}_{c,p}^{\text{chemprot}} \,\|\, \mathrm{d}_{c,(p,z)}^{\text{hci}}\right] \\
f_{\text{outer}} &= W\left(\mathrm{d}_{c,p}^{\text{chemprot}} \otimes \mathrm{d}_{c,(p,z)}^{\text{hci}}\right) \\
\text{gate} &= \sigma\big(W_g\left[\mathrm{d}_{c,(p,z)}^{\text{hci}} \,\|\, \mathrm{d}_{c,p}^{\text{chemprot}}\right]\big) \\
f_{\text{gated-ChemProt}} &= \mathrm{d}_{c,(p,z)}^{\text{hci}} + \text{gate} \odot \mathrm{d}_{c,p}^{\text{chemprot}} \\
f_{\text{gated-HCI}} &= \mathrm{d}_{c,p}^{\text{chemprot}} + \text{gate} \odot \mathrm{d}_{c,(p,z)}^{\text{hci}} \\
f_{\text{gated-weighted}} &= f_{\text{gate}} \odot \mathrm{d}_{c,(p,z)}^{\text{hci}} + (1 - f_{\text{gate}}) \odot \mathrm{d}_{c,p}^{\text{chemprot}} \\
f_{\text{attn-HCI}} &= \mathrm{d}_{c,p}^{\text{chemprot}} + \mathrm{CA}\big(\mathrm{d}_{c,p}^{\text{chemprot}},\, \mathrm{d}_{c,(p,z)}^{\text{hci}}\big) \odot \mathrm{d}_{c,(p,z)}^{\text{hci}} \\
f_{\text{attn-ChemProt}} &= \mathrm{d}_{c,(p,z)}^{\text{hci}} + \mathrm{CA}\big(\mathrm{d}_{c,(p,z)}^{\text{hci}},\, \mathrm{d}_{c,p}^{\text{chemprot}}\big) \odot \mathrm{d}_{c,p}^{\text{chemprot}}
\end{aligned}
\tag{5}
$$

For $f_{\text{attn-HCI}}$, $\mathrm{d}_{c,p}^{\text{chemprot}}$ serves as the base, while $\mathrm{d}_{c,(p,z)}^{\text{hci}}$ is selectively injected. Specifically, $\mathrm{CA}(\mathrm{d}_{c,p}^{\text{chemprot}}, \mathrm{d}_{c,(p,z)}^{\text{hci}})$ computes attention using chemoproteomics as the query over HCI, and the resulting attention weights element-wise modulate $\mathrm{d}_{c,(p,z)}^{\text{hci}}$ before addition. For $f_{\text{attn-ChemProt}}$, the roles are reversed. In Figure 4(c) of the main paper, "Attention" and "Attention (reverse)" correspond to $f_{\text{attn-ChemProt}}$ and $f_{\text{attn-HCI}}$, respectively.

Note, we also explored sharing the cross-modal fusion operations, *i.e.*, common weights in (5) across all tasks. However, this approach reduced overall activity modeling performance, 77.02% to 75.2% mean AUCROC on U2OS and 74.94% to 73.8% mean AUCROC on iNeuron for BiGMINT. Empirically, this can be attributed to (i) diminished per-task adaptation capability, and (ii) conflicting optimization objectives inherent in the multi-task setting.

S.3.4. Biological prior: PPI and TTI

Biological prior is defined by the protein-to-protein interaction (PPI) information derived from STRING [31]. The PPI is computed as the sum of the protein-protein interaction scores based on "experimentally determined interactions" and "database annotations" in STRING. Afterwards, PPI is binarized by including interactions > 0 to produce binarized PPI adjacency matrix $\mathrm{B_P}$.

Task-to-task interaction (TTI) $\mathrm{B_T}$ is derived as, $\mathrm{B_T} \in \{0,1\}^{|\mathcal{T}|\times|\mathcal{T}|}$, as $\mathrm{B_T}(t_{p,z},\ t_{p',z'}) := \mathrm{B_P}(p,p')$ and set $\mathrm{B_T}(t_{p,z}, t_{p,z}) = 0$ to exclude self-interaction. Next,

task-to-task label correlation matrix is computed using the ground truth activity labels in the training set. We threshold B_T at 0.2 and select the top-5 correlated tasks for each task, to retain the most informative task-to-task relations. This sparsifies B_T and promotes information sharing among the most related tasks and prevents excessive or noisy information sharing due to dense or unreliable PPI networks. The aforementioned workflow is presented in Figure 3.

S.4. Experimental setup

S.4.1. Baselines

S.4.1.1. Chemoproteomics Baselines

Chemoproteomics-baselines take the protein sequence and compound SMILES as input, and output continuous binding strength pIC_{50}. We experimented with different methods, namely, PSICHIC, DrugBAN, Bi-directional cross attention between ESM2 protein [22] and RDKit compound [4] embeddings following [16, 36].

We compiled ∼307K compound–protein pIC_{50} pairs spanning ∼85K compounds and 61 proteins. Of these, 20.7% correspond to precise quantitative measurements of protein–compound binding affinity, while the remaining 79.3% represent qualitative constraints indicating whether the true pIC_{50} is greater than or less than a reference value. To jointly leverage both exact and qualitative supervision, we employ a soft-constrained MSE loss that incorporates exact-match, lower-bound, and upper-bound penalties, as defined below. Here, q_i specifies whether the predicted pIC_{50} should be greater than, less than, or equal to the reference value $y_{\mathrm{pIC}_{50},i}$.

$$L_{\text{total}} = \frac{1}{n}\sum_{i=1}^{n} \ell_i$$

$$\ell_i = \begin{cases} (\hat{y}_{\mathrm{pIC}_{50},i} - y_{\mathrm{pIC}_{50},i})^2, & \text{if } q_i \text{ is } =, \\ \left[\mathrm{softplus}\left(\beta\left(y_{\mathrm{pIC}_{50},i} - \hat{y}_{\mathrm{pIC}_{50},i}\right)\right)\right]^2, & \text{if } q_i \text{ is } >, \\ \left[\mathrm{softplus}\left(\beta\left(\hat{y}_{\mathrm{pIC}_{50},i} - y_{\mathrm{pIC}_{50},i}\right)\right)\right]^2, & \text{if } q_i \text{ is } <. \end{cases} \tag{6}$$

where $\hat{y}_{\mathrm{pIC}_{50}}$ is the predicted pIC_{50}, $y_{\mathrm{pIC}_{50}}$ is the annotation, and $\beta > 0$ is the strength of the penalty. When $q_i :=>$, the prediction $\hat{y}_{\mathrm{pIC}_{50}}$ should be larger than the annotation $y_{\mathrm{pIC}_{50}}$. This implies $\left(\beta\left(y_{\mathrm{pIC}_{50},i} - \hat{y}_{\mathrm{pIC}_{50},i}\right)\right) < 0$ and $\left[\mathrm{softplus}\left(\beta\left(y_{\mathrm{pIC}_{50},i} - \hat{y}_{\mathrm{pIC}_{50},i}\right)\right)\right]^2 \to 0$. This approach allows learning from noisy or incomplete supervision without imposing strict equality constraints on the available pIC_{50} measurements.

To obtain the binary compound activity predictions, we then threshold the predicted pIC_{50} against the concentrations $z \in \mathcal{Z}_p$, *i.e.*, c for p at z is active, *i.e.*, $\hat{y}_{z,(p,c)} = 1$, if predicted $\hat{y}_{\mathrm{pIC}_{50},i} > -\log_{10}(z)$.

S.4.1.2. HCI-baselines

HCI-baselines comprises methods using identical HCI FM-derived features as input, but differ in phenotypic feature aggregation mechanisms (mean-pooling, attention-pooling via Multiple Instance Learning (MIL)) and modeling paradigms (multi-label learning (MLL), multi-task learning (MTL)). The baselines are described as follows:

Mean → MLL [13]: It aggregates features across HCI images using mean-pooling. Then, a shared MLP projects the features and a shared classifier across all tasks predicts compound activities. The model is trained with a MLL objective, optimizing a single objective across all tasks.

MIL → MLL [28]: A shared projector maps HCI image features into a common representation, then a MIL aggregator, specifically DSMIL [21], uses attention-pooling to produce an aggregated embedding. Next, a shared classifier is employed to predict compound activities across all tasks.

Mean → MTL [18]: A single shared projector transforms the mean-pooled HCI embedding, followed by task-specific predictors to generate outputs for each task. This multi-task architecture maintains a common intermediate representation while allowing for per-task specialization.

MIL → MTL [28]: A shared projector projects per-image representation, and each task is equipped with a task-specific MIL head that aggregates the projected images representations into a task-specific aggregated HCI embedding. Subsequently, a task-specific predictor maps the aggregated embedding to task labels. This design enables per-task specialization in both aggregation and prediction.

MIL + task-task interaction (TTI) → MTL [28]: This setup builds on MIL → MTL [28], further incorporating the TTI prior (see Section S.3.4) to augment per-task embeddings before processing by task-specific predictors. The TTI integration follows the BiGMINT framework.

S.4.1.3. Multimodal Baselines

Multimodal baselines integrate information from compound (C), protein (P), and HCI (H) using various approaches.

CLOOME → MTL [29]: In this baseline, CLOOME is first used to embed HCI images and compounds into a shared feature space. Matched HCI–compound pairs are pulled closer, while unmatched pairs are pushed apart via contrastive learning. A task-specific predictor trained with an MTL approach then maps the HCI embeddings to the target task labels. CLOOME pretraining follows the protocol described in [29]. For fair comparison, we adopt the same MTL training and validation protocols as BiGMINT.

MolPhenix → MTL [29]: For this baseline, we adopt the inter-sample similarity–aware loss (S2L) proposed by MolPhenix [29] to learn a joint feature space between compound SMILES and HCI images, aligning paired samples via contrastive learning. A task-specific predictor trained with a MTL approach then maps the HCI embeddings to tar-

get task labels, following the same protocol as BiGMINT. Since MolPhenix training data are private, we trained the model on our internal dataset, which uses a fixed HCI treatment dose, unlike the multi-concentration HCI data in [29]. Moreover, to ensure fair comparison, we encoded morphology using our internal HCI foundation model rather than the Phenom1 foundation model used in [29].

MM-Union [25]: This approach is not strictly multimodal. It uses the outputs of the best unimodal baselines, MIL(H) → MTL and PSICHIC, and greedily selects the output from the modality that performs best for each task. It represents the upper-bound performance achievable by combining unimodal methods and only highlights the complementarities of individual modalities.

Concatenate(MIL(H), C) → MTL [32]: This method concatenates the per-task HCI embedding and the compound representation from RDKit, followed by per-task classification. The HCI embeddings are produced using the best performing HCI-baseline, MIL → MTL.

Concatenate(MIL(H), P) → MTL [32]: This method concatenates the per-task HCI embedding and the task-specific protein representation from ESM2, followed by per-task classification. The HCI embeddings are produced using the best performing HCI-baseline, MIL → MTL.

CA(H, CP) → MTL: In this baseline, MIL-based aggregation is replaced by a cross-attention mechanism for phenotypic aggregation across HCI image features. Per-task phenotypic embeddings are generated by a cross-attention module, where chemoproteomics embeddings serve as queries and projected HCI image features act as keys and values. This enables chemoproteomics-guided pooling of HCI image features, enhancing phenotype signals most relevant to the chemoproteomic context. These per-task phenotypic embeddings are then used to predict compound activities in a multi-task setup.

Ablation studies are performed by ablating/swapping different modules of BiGMINT.

Outer(MIL(H), $\mathrm{d}_{c,p}^{\mathbf{chemprot}}$) ~~+ TTI~~ → MTL: TTI prior is excluded, and the chemoproteomics-guided phenotypic aggregators are replaced by MIL aggregators in BiGMINT.

Outer(MIL(H), $\mathrm{d}_{c,p}^{\mathbf{chemprot}}$) + TTI → MTL: Only the chemoproteomics-guided phenotypic aggregations are replaced by MIL aggregators in BiGMINT.

Outer(CA(H, $\mathrm{d}_{c,p}^{\mathbf{chemprot}}$), $\mathrm{d}_{c,p}^{\mathbf{chemprot}}$) ~~+ TTI~~ → MTL: Only the TTI prior is excluded from BiGMINT.

S.4.2. Hyperparameters

List of hyperparameters along with their explored values/ranges for HCI-methods, Chemoproteomics-methods and Multimodal-methods are presented in Table 7.

S.5. Experimental Results

S.5.1. Extended Results and Analysis

1. Table 4 presents the number of high-performing tasks captured across different AUCROC thresholds by BiGMINT, the competing baselines, and the ablations of BiGMINT on U2OS and iNeuron datasets.
2. Table 5 and Figure 4 present the activity modeling performance of different cross-modal fusion operators evaluated within BiGMINT, with TTI excluded to isolate operator effects. Across both cell lines, outer-product fusion achieves the strongest performance, offering an effective and lightweight approach for capturing rich cross-modal correlations, especially under limited annotation.
3. Figure 6 presents the per-task distribution of AUCROC across 5-test folds on U2OS and iNeuron datasets.
4. Figure 7 shows the predictive performance of the best HCI (MIL → MTL), the best Chemproteomics (PSICHIC), and BiGMINT across protein families.
5. Figure 8 presents the effect of biological prior on BiGMINT, evaluated at task-level on U2OS and iNeuron datasets.

S.5.2. Generalization to Unseen Proteins and Assays

We evaluated generalization under biological and assay-level distribution shifts by benchmarking BiGMINT against competing multimodal baselines in an inductive evaluation setting. Specifically, we assessed performance on **40 previously unseen assays**, spanning **25 unseen proteins** and **54 new prediction tasks** across the U2OS and iNeuron datasets, using a linear probing protocol. For each new task, we first identified the most similar task head among the 170 modeled tasks, based on protein similarity derived from PPI relations and assay concentration. Using this task head, we extracted frozen embeddings from the pretrained model and trained a logistic regression classifier using only the labels available for the new task, evaluated via 5-fold cross-validation. As summarized in Table 6, BiGMINT consistently and significantly outperformed all baselines in this inductive setting, demonstrating strong transferability to previously unseen biological and assay spaces.

S.5.3. Computational Cost Analysis

Figure 5 compares model size, peak GPU memory, and average inference latency as the number of tasks increases. Across all metrics, BiGMINT exhibits near-linear scaling with task count, consistent with its task-head–based design. Among variants, BiGMINT with Outer-product fusion incurs higher computational cost than lighter alternatives such as Gated-Add (0.33M vs. 0.07M parameters per task). However, this increased cost remains practical in absolute terms. Even at 1,000 tasks, BiGMINT—including the

	Method	U2OS			iNeuron		
		#Tasks AUCROC $\geq$ 70%	#Tasks AUCROC $\geq$ 80%	#Tasks AUCROC $\geq$ 90%	#Tasks AUCROC $\geq$ 70%	#Tasks AUCROC $\geq$ 80%	#Tasks AUCROC $\geq$ 90%
HCI	Mean $\rightarrow$ MLL [13]	55	18	5	50	20	1
	MIL $\rightarrow$ MLL [28]	64	21	5	47	15	0
	Mean $\rightarrow$ MTL [18]	73	27	8	67	28	4
	MIL $\rightarrow$ MTL [28]	86	33	8	69	31	2
	MIL + TTI $\rightarrow$ MTL [28]	88	34	9	70	34	4
Chemprot	One-hot Logistic Reg. [10]	0	0	0	0	0	0
	Bidirectional CA(C,P) [16]	52	16	3	37	7	2
	DrugBAN [3]	80	31	6	91	32	6
	PSICHIC [19]	103	26	5	111	46	6
Multimodal	MM-Union (MIL(H) $\rightarrow$ MTL, PSICHIC) [25]	125	43	11	120	56	8
	Concatenate (MIL(H), C) $\rightarrow$ MTL [32]	111	44	9	96	39	8
	Concatenate (MIL(H), P) $\rightarrow$ MTL [32]	104	43	8	108	47	9
	CA(H, $d_{c,p}^{\text{chemprot}}$) $\rightarrow$ MTL	101	41	10	83	32	4
	BiGMINT: Outer(CA(H, $d_{c,p}^{\text{chemprot}}$), $d_{c,p}^{\text{chemprot}}$) + TTI $\rightarrow$ MTL	**147**	**67**	**19**	**127**	**59**	**13**
	Ablating BiGMINT						
	Outer(CA(H, $d_{c,p}^{\text{chemprot}}$), $d_{c,p}^{\text{chemprot}}$) ~~+ TTI~~ $\rightarrow$ MTL	136	60	16	116	49	9
	Outer(MIL(H), $d_{c,p}^{\text{chemprot}}$) + TTI $\rightarrow$ MTL	142	61	17	124	52	11
	Outer(MIL(H), $d_{c,p}^{\text{chemprot}}$) ~~+ TTI~~ $\rightarrow$ MTL	130	54	14	115	48	9

Table 4. Number of high-performing tasks achieving AUCROC $\geq$ 70%, 80%, and 90% thresholds on U2OS and iNeuron datasets. Best results in each block are underlined and the overall best are in **bold**. Notations: H: HCI features, C: Compound features, P: Protein features, $d_{c,p}^{\text{chemprot}}$: Chemprotomics features, CA: CrossAttention.

	Method	U2OS			iNeuron		
		AUCROC	AUPRC	Macro F1	AUCROC	AUPRC	Macro F1
MIL$\rightarrow$ MTL	Concatenation	76.05	54.76	63.31	73.23	53.23	61.34
	Gated add	74.39	52.72	62.17	71.37	51.02	60.27
	Gated add inverse	74.83	53.05	62.48	72.33	52.08	61.03
	Gated add weighted	73.72	51.46	62.11	70.49	50.23	60.08
	Gated e-add	74.84	52.88	62.61	72.29	52.19	60.99
	Gated e-add inverse	75.91	54.20	63.33	71.76	51.79	60.28
	Gated e-add weighted	75.50	53.99	62.92	72.32	52.05	61.20
	Attention	75.68	54.11	62.82	72.93	52.85	61.28
	Attention inverse	75.44	54.04	62.90	72.92	53.01	61.37
	Outer product	**76.41**	**54.90**	**63.76**	**74.66**	**55.15**	**62.13**
CA$\rightarrow$ MTL	Concatenation	76.36	54.92	63.61	73.29	53.33	61.22
	Gated add	75.08	52.86	62.67	71.35	51.40	60.29
	Gated add inverse	75.21	53.66	62.82	72.30	52.37	61.03
	Gated add weighted	74.80	52.59	62.33	71.39	51.36	60.62
	Gated e-add	75.22	52.97	62.60	72.29	52.33	61.11
	Gated e-add inverse	75.39	54.03	63.12	72.57	52.58	61.38
	Gated e-add weighted	75.88	54.72	63.71	72.46	52.56	61.19
	Attention	76.18	54.87	63.35	73.30	53.25	61.75
	Attention inverse	76.22	54.88	63.54	73.15	53.61	61.25
	Outer product	**77.02**	**55.90**	**64.38**	**74.94**	**56.18**	**63.18**

Table 5. Comparison of phenotype aggregators (MIL vs. CrossAttention using chemoproteomics-guidance) and cross-modal fusion operators on U2OS and iNeuron datasets for multi-task compound activity modeling.

Method	U2OS				iNeuron			
	AUCROC	AUCPRC	Macro F1	AUCROC $\geq$80%	AUCROC	AUCPRC	Macro F1	AUCROC $\geq$80%
Concatenate (MIL(H), C) $\rightarrow$ MTL	79.60	58.48	66.53	27 (-30%)	77.75	60.67	64.27	23 (-39%)
Concatenate (MIL(H), P) $\rightarrow$ MTL	78.47	57.40	65.56	26 (-35%)	78.62	61.10	65.43	24 (-33%)
CA(H, $d_{c,p}^{\text{chemprot}}$) $\rightarrow$ MTL	77.60	56.13	65.54	24 (-46%)	75.71	56.61	62.75	19 (-68%)
BiGMINT: Outer(CA(H, $d_{c,p}^{\text{chemprot}}$), $d_{c,p}^{\text{chemprot}}$) + TTI $\rightarrow$ MTL	**83.34**	**62.73**	**69.97**	**35** ⚓	**82.76**	**64.80**	**69.80**	**32** ⚓

Table 6. Generalization performance of the multimodal methods evaluated under inductive biological shifts on U2OS and iNeuron datasets.

Outer-product variant—operates within modest resource requirements, with peak GPU memory below <3GiB and inference latency <1s. This additional cost is justified by clear performance gains, including improvements in mean AUCROC (77.02 vs. 75.08) and a 25% increase in high-performing tasks (60 vs. 48 for AUCROC$\geq$80%), as shown in Fig. 4. In practical drug discovery settings, these accuracy gains substantially outweigh the marginal

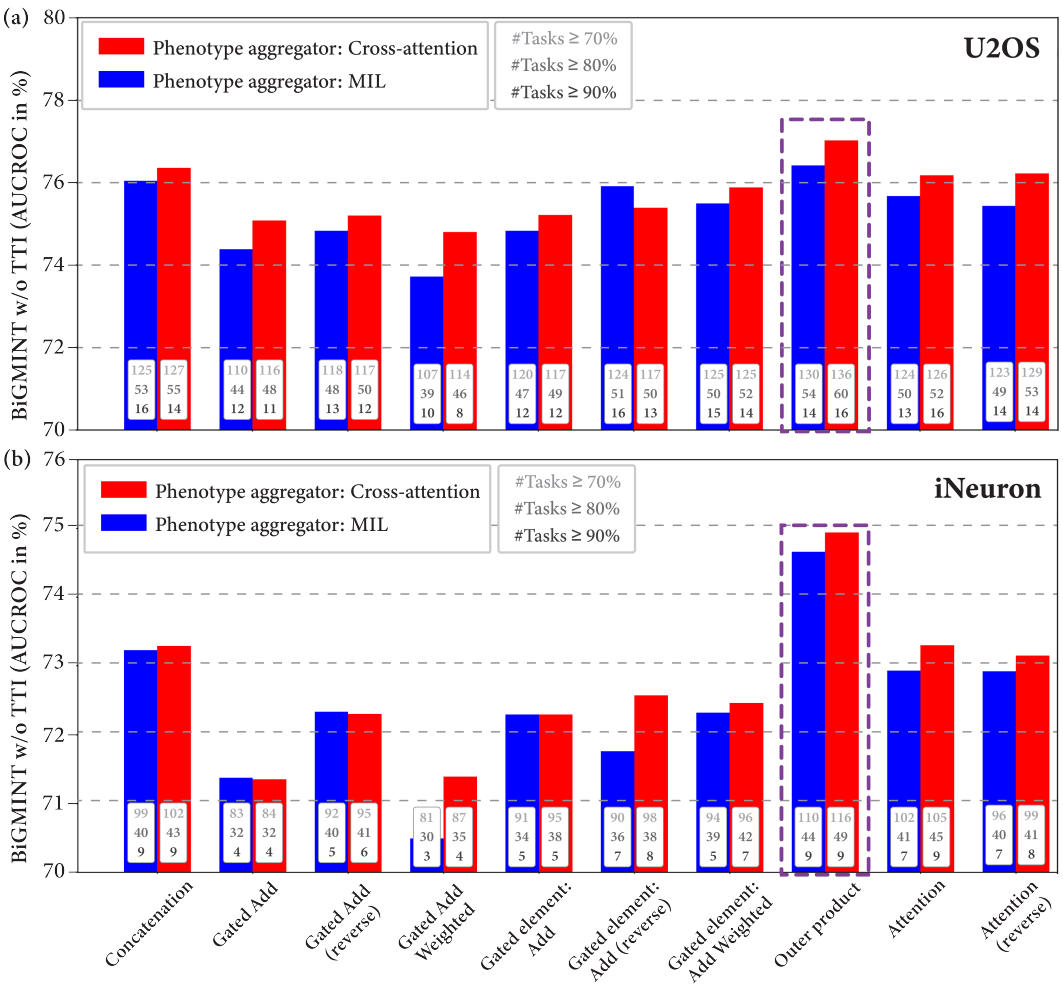

Figure 4. Ablation study of phenotype aggregators and cross-modal fusion strategies on U2OS and iNeuron datasets. The cross-attention based phenotypic aggregation consistently produced better mean AUCROC and high-performing task-coverage across all cross-modal operators on both datasets. Among the cross-modal fusion operators, outer-product consistently produced the best performance.

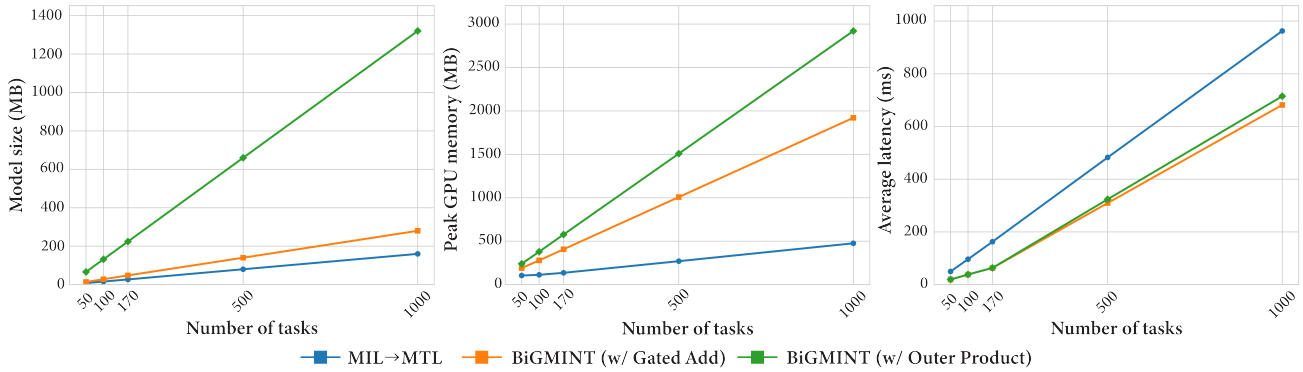

Figure 5. Scaling of memory footprint and inference latency with task count.

deployment overhead. For resource-constrained scenarios, lower-cost fusion strategies (*e.g.*, Gated-Add) provide an effective efficiency–accuracy trade-off while preserving favorable scaling behavior.

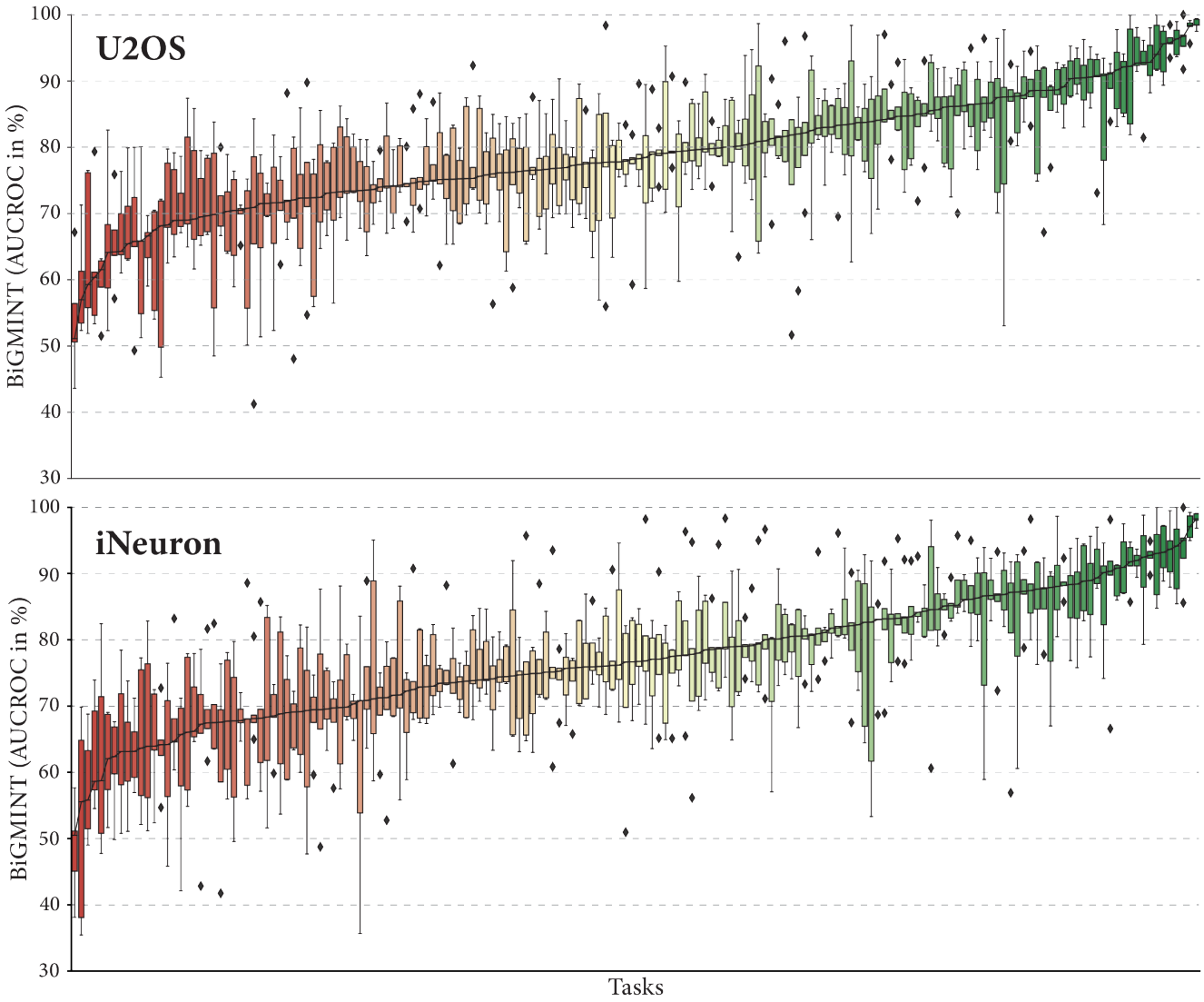

Figure 6. Per-task distribution of AUCROC across 5-test folds on U2OS and iNeuron datasets. The tasks are sorted by median AUCROC scores. The range of median scores indicate the diversity of task complexities. Further, the AUCROC variations are observed to be large for certain tasks, which can be attributed to the underlying imbalanced class distribution.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González. Gated multimodal networks. *Neural Computing and Applications*, 32(14):10209–10228, 2020. 4

[2] John Arevalo, Ellen Su, Jessica D Ewald, Robert van Dijk, Anne E Carpenter, and Shantanu Singh. Evaluating batch correction methods for image-based cell profiling. *Nature Communications*, 15(1):6516, 2024. 1

[3] Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023. 8

[4] A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12(1):51, 2020. 3, 6

[5] Olivier JM Bequignon, Brandon J Bongers, Willem Jespers, Adriaan P IJzerman, B van der Water, and Gerard JP van Westen. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *Journal of cheminformatics*, 15(1): 3, 2023. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 4

[7] Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric

(a) **HCI methods**

Hyperparameter	Values / Ranges
Epochs	100
Early stopping	10
Patience	10
Batch size	128
Optimizer	AdamW
Learning rates	[0.0001, 0.00005]
Weight decay	0.0005

(b) **Chemoproteomics methods**

Hyperparameter	Values / Ranges
Epochs	30
Gradient clipping	[0.1, 0.0]
Batch size	512
Optimizer	AdamW
Learning rates	[0.001, 0.002, 0.004, 0.008]
Weight decay	0.01
Penalty strength β	[0.1, 0.5]

(c) **Multimodal methods**

Hyperparameter	Values / Ranges
Epochs	100
Early stopping	10
Patience	10
Batch size	512
Optimizer	AdamW
Learning rates	[0.0001, 0.00005]
Weight decay	0.0005

Table 7. Hyperparameters and explored values/ranges for different methods.

Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, pages 2023–03, 2023. 2, 3

[8] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020. 4

[9] Tingting Chen, Xinjun Ma, Xingde Ying, Wenzhe Wang, Chunnv Yuan, Weiguo Lu, Danny Z Chen, and Jian Wu. Multi-modal fusion learning for cervical dysplasia diagnosis. In *IEEE ISBI*, pages 1505–1509, 2019. 4

[10] Tianyu Cui, Song-Jun Xu, Artem Moskalev, Shuwei Li, Tommaso Mansi, Mangal Prakash, and Rui Liao. Infosem: A deep generative model with informative priors for gene regulatory network inference. In *International Conference on Machine Learning*, 2025. 8

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[12] Jan P Engelmann, Alessandro Palma, Jakub M Tomczak, Fabian Theis, and Francesco Paolo Casale. Mixed models with multiple instance learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3664–3672, 2024. 1

[13] Johan Fredin Haslum, Charles-Hugues Lardeau, Johan Karlsson, Riku Turkki, Karl-Johan Leuchowius, Kevin Smith, and Erik Müllers. Cell painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. *Nature Communications*, 15(1):3470, 2024. 1, 6, 8

[14] Yulu Guan, Hui Cui, Yiyue Xu, Qiangguo Jin, Tian Feng, Huawei Tu, Ping Xuan, Wanlong Li, Linlin Wang, and Been-Lirn Duh. Predicting esophageal fistula risks using a multimodal self-attention network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–730, 2021. 4

[15] Dorota Herman, Maciej M Kanduła, Lorena GA Freitas, Caressa van Dongen, Thanh Le Van, Natalie Mesens, Steffen Jaensch, Emmanuel Gustin, Liesbeth Micholt, Charles-Hugues Lardeau, et al. Leveraging cell painting images to expand the applicability domain and actively improve deep learning quantitative structure–activity relationship models. *Chemical Research in Toxicology*, 36(7):1028–1036, 2023. 1, 3

[16] Markus Hiller, Krista A Ehinger, and Tom Drummond. Perceiving longer sequences with bi-directional cross-attention transformers. In *Advances in Neural Information Processing Systems*, pages 94097–94129, 2024. 6, 8

[17] Markus Hofmarcher, Elisabeth Rumetshofer, Djork-Arne Clevert, Sepp Hochreiter, and Gunter Klambauer. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of Chemical Information and Modeling*, 59(3):1163–1171, 2019. 1

[18] Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *bioRxiv*, 2023. 6, 8

[19] Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 6(6):673–687, 2024. 3, 8

[20] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768, 2024. 2

[21] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 14318–14328, 2021. 6

[22] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. 3, 6

[23] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007. 2

[24] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015. 2

[25] Nikita Moshkov, Tim Becker, Kevin Yang, Peter Horvath, Vlado Dancik, Bridget K Wagner, Paul A Clemons, Shantanu Singh, Anne E Carpenter, and Juan C Caicedo. Predicting compound activity from phenotypic profiles and chemical structures. *Nature communications*, 14(1):1967, 2023. 3, 7, 8

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 4

[27] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025. 2

[28] Pushpak Pati, Hsiu-Chi Cheng, Steffen Jaensch, Walid M Abdelmoula, Krishna Chaitanya, Michiel Van Dyck, Tomé Albuquerque, Samantha Allen, Litao Zhang, Tommaso Mansi, et al. Momil: Mixture of multi-instance learners for modeling multiple compound activities in high content imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381, 2025. 6, 8

[29] Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023. 6, 7

[30] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, et al. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of cheminformatics*, 9(1):17, 2017. 3

[31] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. 5

[32] Guangyan Tian, Philip J Harrison, Akshai P Sreenivasan, Jordi Carreras-Puigvert, and Ola Spjuth. Combining molecular and cell painting image data for mechanism of action prediction. *Artificial Intelligence in the Life Sciences*, 3:100060, 2023. 7, 8

[33] Zhiqin Wang, Ruiqing Li, Minghui Wang, and Ao Li. Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*, 37(18):2963–2970, 2021. 4

[34] Erin Weisbart, Ankur Kumar, John Arevalo, Anne E Carpenter, Beth A Cimini, and Shantanu Singh. Cell painting gallery: an open resource for image-based profiling. *Nature Methods*, 21(10):1775–1777, 2024. 3

[35] Christopher Wolff, Martin Neuenschwander, Carsten Jörn Beese, Divya Sitani, Maria C Ramos, Alzbeta Srovnalova, María José Varela, Pavel Polishchuk, Katholiki E Skopelitou, Ctibor Škuta, et al. Morphological profiling data resource enables prediction of chemical compound properties. *Iscience*, 28(5), 2025. 2

[36] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv:2111.07783*, 2021. 6

[37] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024. 2

[38] Jun Zhou, Kai Chen, Linlin Xu, Qi Dou, and Jing Qin. Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13967–13977, 2023. 4

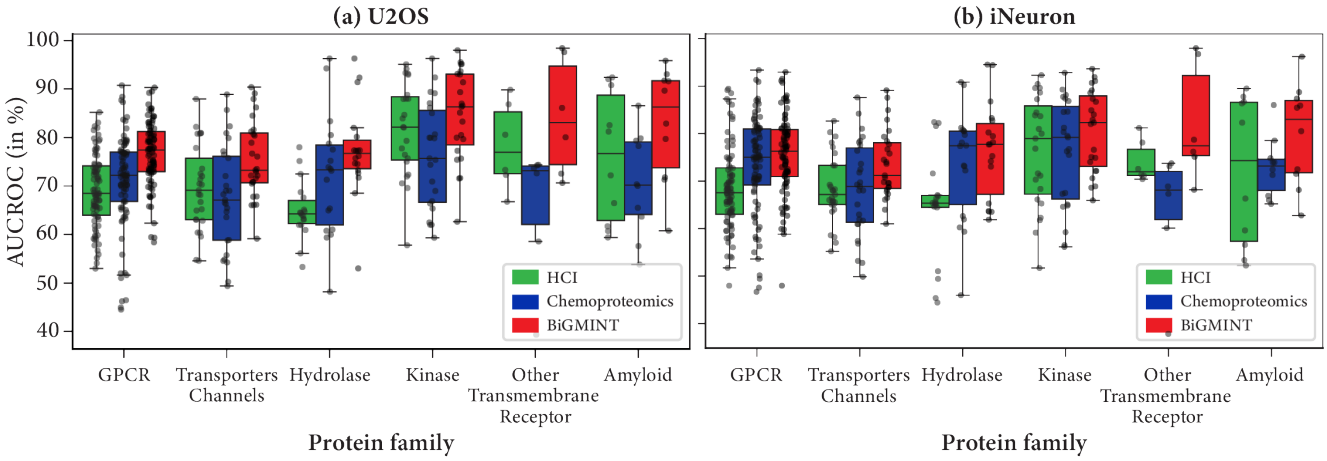

Figure 7. Predictive performance of the best HCI (MIL → MTL), the best Chemproteomics (PSICHIC), and BiGMINT methods within each protein family. BiGMINT produced consistently higher median than competing methods.

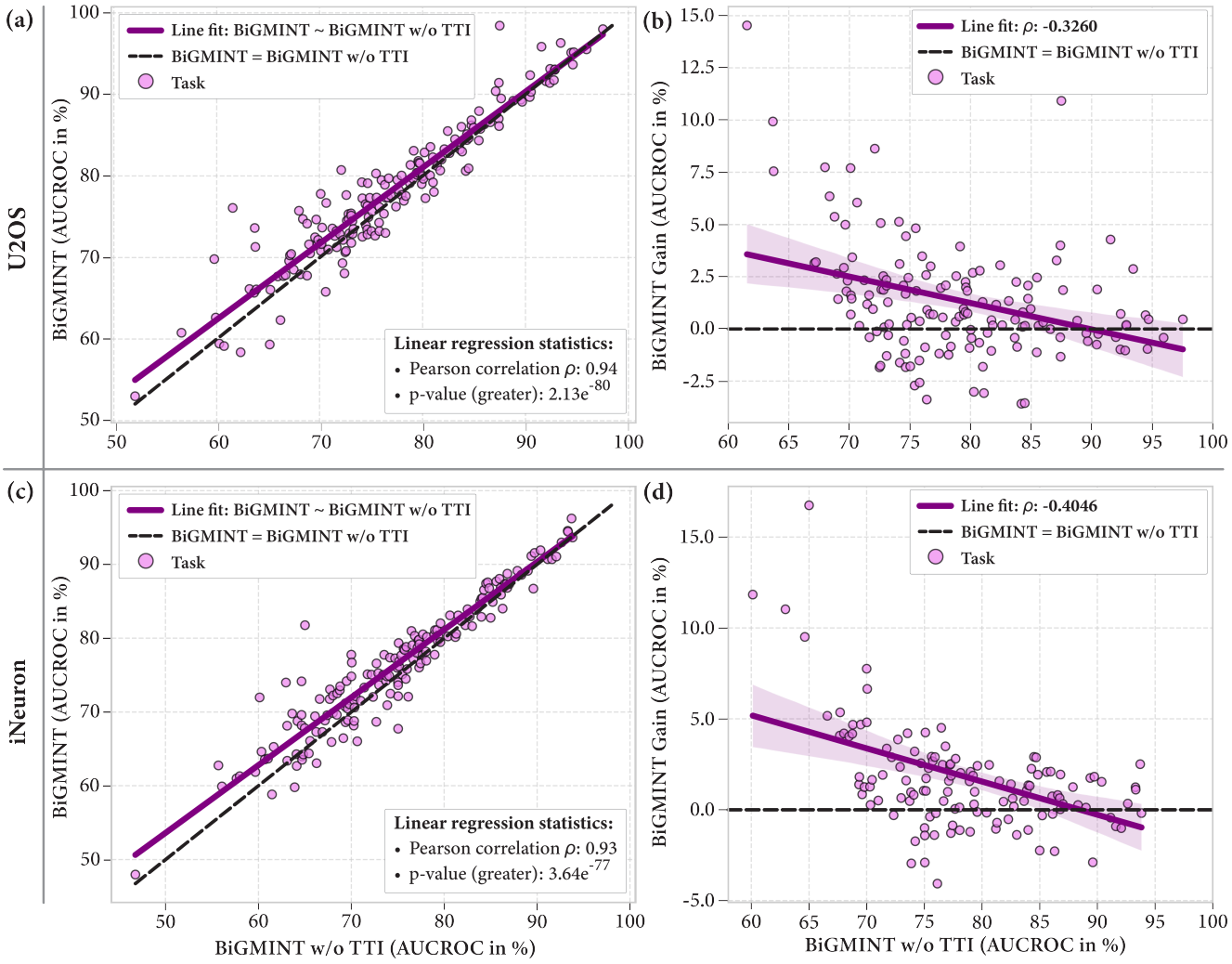

Figure 8. Effect of biological prior on BiGMINT, evaluated at task-level. (a, c) Trend between BiGMINT with and without TTI on U2OS and iNeuron datasets. (b, d) Trend of relative gain with respect to baseline BiGMINT without TTI on U2OS and iNeuron datasets.