FROM REGRESSION TO DOSE–RESPONSE: PREDICTING GPCR ACTIVITY AND EC_{50} VIA CONCENTRATION-CONDITIONAL MODELLING

Anonymous authors
Paper under double-blind review

ABSTRACT

ML models have revolutionised drug discovery, yet they still struggle with modelling the molecule-induced response of G-protein coupled receptors (GPCRs) — cellular membrane "sensors" that account for nearly 30% targets of approved drugs. The central challenge in GPCRs lies in predicting activity and potency (i.e., half maximal effective concentration, EC_{50}). To estimate these properties, we propose a unified learning framework inspired by *in vitro* dose-response assays that models concentration-conditional activation. Specifically, for each training protein-molecule pair (s, m), we uniformly sample concentrations c and generate labels that are deterministic far from the EC_{50} and soft in the transition region. We then train a concentration-conditional model that learns the probability $P(active|s, m, c)$ at any c. Finally, querying the model across concentrations enables fitting a Hill curve from which both activity (upper asymptote) and EC_{50} (inflection point in $\log c$) are derived. On the challenging M2OR dataset, our approach achieves an EC_{50} estimation error of 0.683 log units, an improvement of 0.399 log units compared to drug–target affinity baselines, 0.205 log units compared to censored regression, and 0.433 log units compared to the Boltz-2 affinity module, while achieving comparable performance in activity classification to the SOTA baselines (MCC 0.665 in the i.i.d. setting).

1 INTRODUCTION

Machine learning models in protein biology (Jumper et al., 2021; Watson et al., 2023; Passaro et al., 2025) have rapidly advanced the field, substantially accelerating drug discovery pipelines (Du et al., 2024). Despite these advances, modelling the molecule-induced response of G-protein coupled receptors (GPCRs) — which account for around 30% of approved drug targets (Sriram & Insel, 2018) — remains notoriously difficult. A single point mutation can alter receptor response (de March et al., 2018), and subtle differences in ligand structure can significantly change protein-ligand interactions, in some cases even converting an active compound into an inhibitor (Kosar et al., 2024; Qin et al., 2022; Heydenreich et al., 2023).

The largest subfamily of GPCRs are olfactory receptors (ORs), which constitute 49% of all genes encoding GPCRs (Bjarnadóttir et al., 2006; Niimura & Nei, 2003). The pattern of activated ORs encodes odour identity (Malnic et al., 1999; Nara et al., 2011), where ORs with lower EC_{50} values appear to have a major impact on the odour code (Junek et al., 2010; Spors & Grinvald, 2002; Wilson et al., 2017). Beyond their role in olfaction, OR transcripts have been found in various tissues, including heart, kidney, liver, lungs, prostate, brain, or leukaemia cells, and hold a promise for potential therapeutic and diagnostic applications in diseases such as asthma, obesity, diabetes, and cancer (Lee et al., 2019).

The key properties of interest for GPCR drug discovery are (i) binary activity (i.e., whether a molecule elicits any measurable response) and (ii) the concentration required to reach 50% of the maximal response (EC_{50}). Currently, the method of choice to estimate these properties is *in vitro* experiments, which usually consist of screening rounds followed by a detailed dose-response assay that allows to estimate EC_{50} for a candidate protein-molecule pair. However, despite the progress in the throughput of *in vitro* assays driven by engineered heterologous systems (Saito et al., 2004; 2009), the most

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

comprehensive database of *in vitro* OR-molecule experiments, M2OR (Lalis et al., 2024a), lists 6157 dose-response assays, of which only 1663 observe a cellular response and estimate EC_{50}. Thus, scalable *in silico* models capable of predicting activity and EC_{50} are essential for characterising odour coding and assessing the drugability of ORs.

In this work, we draw an analogy between experimental dose-response assays and probabilistic learning. In *in vitro* experiments, responses are measured at multiple concentrations, activity is determined by the maximal response, and EC_{50} is estimated by fitting a Hill curve. Motivated by this process, we propose a learning framework that models the conditional probability of activation $P(active|s, m, c)$ for a protein-molecule pair (s, m) at concentration c. The model is trained using supervision derived from experimental EC_{50} measurements, which avoids reliance on raw assay readouts that are not directly comparable across experimental settings and are not available in the public datasets. At inference time, we query the model across a range of concentrations — mimicking a virtual dose-response assay — and fit a logistic Hill curve to the predicted probabilities. This procedure yields both a final activity decision (the curve's maximum) and EC_{50} (the curve's inflection point). In summary, this work makes the following main contributions:

- We introduce a concentration-conditional framework that recasts EC_{50} estimation as a virtual dose-response classification task and jointly recovers both activity and potency.
- The framework naturally accommodates inactive pairs and censored lower-bound observations through concentration-dependent supervision.
- Extensive validation shows that this approach improves EC_{50} estimation on the challenging M2OR dataset, and the improvement is robust across out-of-distribution splits, alternative estimators, decision thresholds, and concentration processing in the model architecture.

2 Related work

Despite the potential of ORs as therapeutic targets and their key role in mammalian olfaction, there is only a limited number of models designed to predict OR-molecule activity, and no previous approach addresses the challenge of predicting EC_{50}. The earliest studies employed SVMs (Kowalewski & Ray, 2020) and random forests (Cong et al., 2022) to predict the responses of limited subsets of ORs and molecules with sufficient *in vitro* data. Gupta et al. (Gupta et al., 2021) used BiLSTM (Graves & Schmidhuber, 2005) to predict the OR-molecule activity based on SMILES and the receptor's primary sequence. The current state-of-the-art (Hladiš et al., 2023) represents the molecular topology as a graph and copies the receptor's [CLS] token from ProtBERT (Elnaggar et al., 2021) to all nodes in the graph. The model then employs a tailored graph neural network to predict the probability that a molecule activates an OR. Recently, MAARDTI (Zhan et al., 2025) has been applied to predict OR activation. Although competitive, the reported performance falls short of the current SOTA, highlighting the challenging nature of OR-molecule activity prediction.

Dose–response modelling has also been investigated for drug–cell inhibitory effects (Alonso Campana et al., 2024). In that setting, the authors assume access to experimental readouts at each concentration and explicitly fit the entire dose–response curve. In contrast, our framework operates in a substantially more restricted regime, where the curve is unknown and only the inflection point of the 4-parameter logistic model (i.e., the EC_{50}) or the information that no response is observed is available.

3 Dose-response curve

The relationship between the concentration of a compound and the receptor-mediated response of a cell is typically assessed using functional assays, which are described by dose-response curves. For concentration c on a logarithmic scale, the dose-response curve can be characterised by a generalised logistic model, or Hill curve, (Neubig et al., 2003):

$$g(c) = \frac{M - b}{1 + 10^{-q(c - \mathrm{EC}_{50})}} + b \tag{1}$$

where M is the efficacy, b is the basal activity, q is the slope, and EC_{50} is the 50% effective concentration and the inflection point of the curve. To estimate the parameters of the curve, several concentrations of the compound over several orders of magnitude are tested *in vitro* (Mainland et al.,

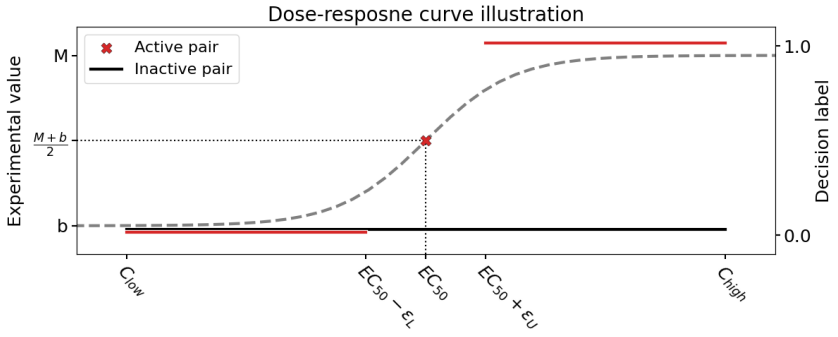


Figure 1: Example of a dose-response curve for an active (grey/red) and inactive (black) pair. The grey line corresponds to the true active curve, which is unknown except for the EC_{50} value (red cross). The red and black lines are concentration intervals corresponding to the training labels on the right y-axis for active and inactive pairs, respectively. The efficacy M, basal activity b, and half response $\frac{M+b}{2}$ are experimentally measured and not available in the data. In Algorithm 1, a soft label $L \sim \text{Unif}(0, 1)$ is sampled for concentrations $c \in (\text{EC}_{50} - \epsilon_L, \text{EC}_{50} + \epsilon_U)$.

2014; Saito et al., 2009). However, due to different readouts depending on the experimental settings (e.g., luciferase, Ca^{2+}, etc.), raw dose-response curve data are not directly comparable, and most data sources publish only EC_{50} values or information that there is no response and the curve is flat. Formally, we set $\text{EC}_{50} = +\infty$ for the flat "inactive" curves.

4 EC_{50} PREDICTION AS CONDITIONAL RESPONSE LEARNING

A widely used approach to train a model to predict EC_{50} is to use l_2 loss and estimate the values as a regression. However, this requires treating inactive pairs, and the method of choice is to assign them an arbitrary large-enough value (Huang et al., 2021; Lu et al., 2025; Shah et al., 2025; Kroll et al., 2024). In contrast, we propose an alternative strategy that formulates EC_{50} prediction as a concentration-conditioned classification task. Similarly to how dose-response curves are obtained from *in vitro* experiments, we model the protein-molecule response at a given molecular concentration in a fixed concentration range, and estimate EC_{50} by fitting (1) to predictions at concentrations spanning several orders of magnitude.

We assume that the model has access to EC_{50}, the other parameters of the curve (1) are unknown, and raw readout data are not available. However, EC_{50} provides a way to draw surrogate decision samples from a normalised dose-response curve $f(c) = \frac{g(c)-b}{M-b}$. The monotonicity of the curve implies that for low concentrations $c \ll \text{EC}_{50}$ the protein-molecule pair would be considered inactive *in vitro*, whereas for high concentrations $c \gg \text{EC}_{50}$ the pair would be active (Figure 1). Therefore, given a dataset $\mathcal{D} = \{(s^i, m^i), \text{EC}_{50}^i\}_i$ of EC_{50} values for each protein-molecule pair (s^i, m^i), we can construct binary training data $\mathcal{B} = \{(s^i, m^i, c^{i,j}), L^{i,j}\}_{i,j}$ where for each concentration $c^{i,j} \sim \text{Unif}(C_{low}, C_{high})$ we sample a decision $L^{i,j} \in \{0, 1\}$ whether the pair would be considered active. Then we train a model on $\mathcal{B}$, that predicts the probability of activation at a given concentration $P(active|s, m, c)$. Finally, we estimate EC_{50} of a pair (s^i, m^i) by fitting (1) to predictions $\{P(active|s^i, m^i, c^{i,j})\}_j$ at several concentrations, similarly to how the dose-response curve is fitted from *in vitro* experiments.

By definition, the probability of activation at $c = \text{EC}_{50}$ corresponds to 0.5, and we uniformly sample $L^{i,j} \sim \text{Unif}(0, 1)$ for concentrations around EC_{50} based on margins ϵ_L and ϵ_U. See detailed training and inference procedures in Algorithm 1 and Algorithm 2. Inherently, there is an imbalance between the number of active and inactive pairs, and this imbalance is changed in each batch due to the sampling. Thus, we derive dynamic sample weights calculated for each batch, with the details given in Section B.

The above strategy reformulates the traditional regression problem as a concentration-conditioned classification task that draws samples from a surrogate of a normalised dose-response curve. Al-

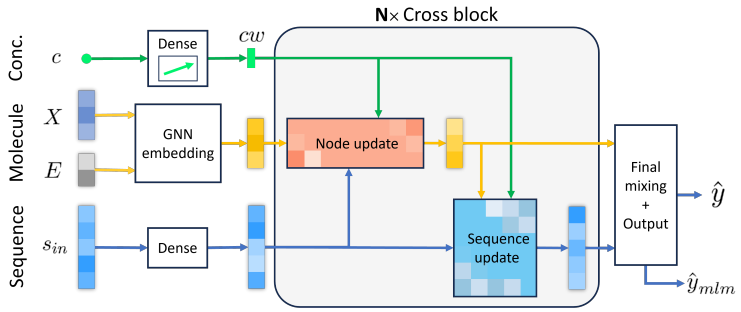


Figure 2: Outline of ASMI-DR architecture. The inputs to the model are a molecular graph m with node features X and edge features E, molecular concentration c, and sequence embedding s_{in}. The molecular graph is passed through the GNN embedding layer and then processed together with the sequence representation and concentration in a series of cross blocks. In each cross block, described in Figure D2, the concentration-dependent interaction between the molecule and the protein is modelled by iteratively updating node and sequence embeddings. The updated representations are then passed through multi-head attention in the final mixing layer and finally two output heads for response prediction and masked language modelling (MLM).

though M and b are assumed to be unknown and generally different for each protein-molecule pair, normalisation $f(c) = \frac{g(c)-b}{M-b}$ only rescales the y-axis, and the inflection point (i.e., EC_{50}) remains the same when the curve is fitted during inference in Algorithm 2.

Beyond active pairs with a reported EC_{50}, a dose–response assay may show an increased response, but the EC_{50} lies outside the tested concentration range. Therefore, the curve cannot be fitted, and only the lower bound c_t on the EC_{50} is available in the data. In such cases, negative samples can still be drawn for $c < c_t$ and we provide further details on this case in Section C.

5 Model

The concentration-conditional formulation introduced in Section 4 is architecture-agnostic and can be combined with any model that represents a protein, molecule, and concentration. For our experiments, we instantiate it with ASMI-DR (*Attention-based Sequence–Molecule Interaction for Dose-Response prediction*), which maps a protein sequence s, molecular graph m, and concentration c to a score $f_\theta(s, m, c)$. The architecture is illustrated in Figure 2 and full details are provided in Section D.

Protein sequences are represented using frozen ESM-2 embeddings (Lin et al., 2023) and molecules are represented as graphs with standard atom and bond features. The molecular and protein representations are processed jointly through a sequence of cross-attention blocks, allowing atom representations to attend to amino acids and amino-acid representations to attend to molecular atoms.

To make these interactions concentration dependent, the scalar concentration c is mapped to a learned embedding c_w, which is concatenated to the queries of the cross-attention layers. Thus, the same protein–molecule pair can produce different interaction representations at different concentrations. After the cross-attention blocks, the protein and molecule embeddings are combined by a final self-attention layer, pooled, and passed to an MLP output head.

We note that the precise mechanism used to incorporate concentration is not essential to the proposed learning formulation. In Section F.2, we also evaluate a simpler variant in which concentration is concatenated only before the output MLP. This variant remains competitive, while concentration-aware cross-attention provides an additional improvement in EC_{50} estimation.

6 Experiments

We evaluate whether concentration-conditioned learning improves EC_{50} estimation while retaining accurate activity prediction. We first compare against regression, censored regression, classification, and structure-based baselines on M2OR (Lalis et al., 2024a) under i.i.d. and OOD splits. We then test

whether the observed gains are robust to the activity threshold, the EC_{50} estimator, and the model's treatment of the concentration. Finally, we evaluate transfer to standard binding-affinity benchmarks.

6.1 Experimental setup

We primarily evaluate our framework on M2OR v1.2.0 (Lalis et al., 2024a), which is a curated collection of functional olfactory receptor-molecule assays. The full database contains 77,611 experiments spanning 1,402 protein sequences and 771 molecules. After preprocessing (Section H), it contains 5,773 dose-response samples, including 1,427 active pairs with a measured EC_{50} and 4,346 inactive pairs. M2OR additionally lists 60,256 primary and secondary screening measurements, which we use only for the screening comparison. To further assess whether the concentration-conditional formulation generalises beyond GPCRs, we also evaluate it on the standard drug–target affinity datasets DAVIS (Davis et al., 2011) and BindingDB (Gilson et al., 2015; Huang et al., 2021), where the target quantity is the dissociation constant K_d. After preprocessing, the two datasets contain 24,548 and 42,234 measured K_d values, respectively.

On M2OR we evaluate a random i.i.d. and four OOD settings following Hladiš et al. (2023): *Sequence-Single* and *Molecule-Single* splits hold out previously unseen individual receptors or molecules, while *Sequence-Cluster* and *Molecule-Cluster* hold out entire sequence or molecular clusters, respectively. The latter two therefore measure extrapolation to more distant protein families and chemical scaffolds. Each experiment uses five splits with approximately 20% test data. Full construction details are in Section H.1.1.

To run the inference in Algorithm 2, we predict the responses of the protein-molecule pairs at molecular concentrations spanning the range 10^{-8}M to 10^{-2}M for M2OR and 10^{-12}M to 10^{-4}M for DAVIS and BindingDB with a step of 0.1 log units. This covers 98.5% of all available EC_{50} values in M2OR and 100% and 99.9% of K_d values in DAVIS and BindingDB, respectively. We exclude values outside of these ranges from test sets.

We evaluate activity prediction using Matthews correlation coefficient (MCC), which is appropriate for the imbalanced activity labels, and report precision and recall in the appendix. EC_{50} estimation is evaluated using root mean squared log error (RMSLE) and Spearman's ρ. Because EC_{50} is defined only for experimentally active pairs, the primary paired comparison computes EC_{50} metrics on the intersection of true positive activity predictions of ASMI-DR and DTA baselines. To verify that conclusions do not depend on this paired subset, Section J.2 and J.3 additionally report each model on its own true-positive predictions, and Section 6.3 examines the recall-RMSLE trade-off across activity thresholds. We additionally report the estimated in vitro EC_{50} measurement error as a reference lower bound.

Baselines. We compare against three groups of baselines. Regression baselines include DTIAM (Lu et al., 2025), DeepDTAGen (Shah et al., 2025), and ProSmith (Kroll et al., 2024), together with ASMI-Reg, which uses the same backbone as ASMI-DR but directly regresses EC_{50} without concentration as an input. We follow standard practice (Huang et al., 2021; Shah et al., 2025; Lu et al., 2025; Kroll et al., 2024) and we set EC_{50} and K_d of inactive pairs to 1M and 10μM, respectively. To isolate the effect of treating inactive observations correctly, we additionally include ASMI-Cen, the same backbone trained with censored regression. Activity-classification baselines include HyperAttentionDTI (Zhao et al., 2021), MolTrans (Huang et al., 2020), MAARDTI (Zhan et al., 2025), GNN-CLS (Hladiš et al., 2023), and the concentration-free ASMI-Prob. To take the tertiary structure into account, structure-based baselines include Boltz-2 (Passaro et al., 2025) and molecular docking performed on 868 (300 active and 568 inactive) pairs corresponding to the OR sequences available at AlphaFold DB (Varadi et al., 2023; 2021). We additionally compare activity predictions to primary and secondary experimental screening outcomes. Full training and evaluation details are given in Section I.

6.2 Concentration-conditioned learning improves EC_{50} estimation

As can be seen in Figure 3, ASMI-DR trained on the concentration-conditional task achieves substantially lower EC_{50} error than the evaluated DTA regression baselines. In the i.i.d. setting, it reaches an RMSLE of 0.683, improving over the strongest DTA baseline by 0.399 log units. The same advantage persists across all four OOD settings, including splits that hold out previously unseen

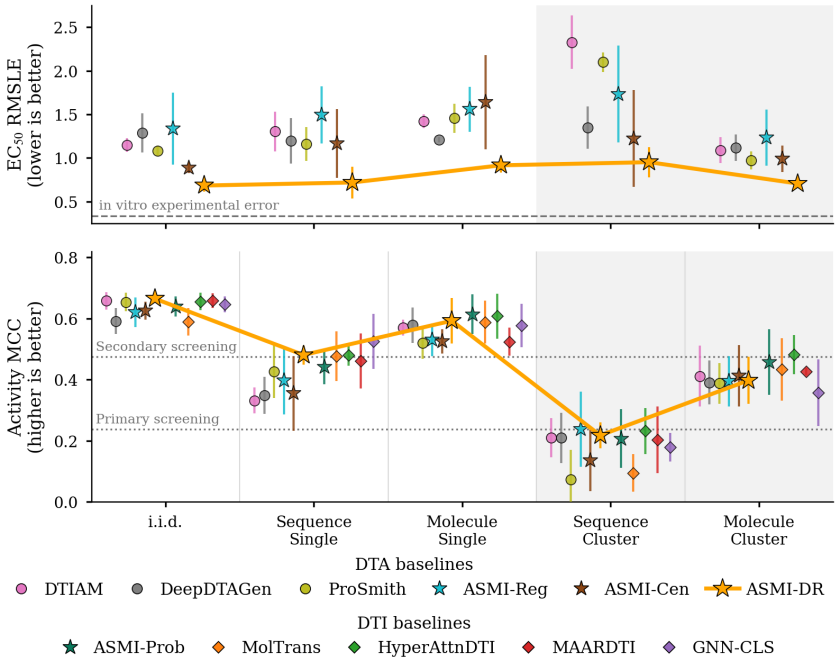


Figure 3: EC_{50} estimation error (top) and activity classification (bottom) across the i.i.d. and four OOD evaluation scenarios of M2OR. RMSLE is computed on the intersection of the correctly predicted active pairs of ASMI-DR and the DTA baselines, so all potency values within a scenario are evaluated on the same pairs. Error bars denote one standard deviation across five splits. The line connecting ASMI-DR across scenarios is a visual guide and intermediate positions have no meaning.

receptor sequences, molecules, sequence clusters, or molecular scaffolds. ASMI-DR also outperforms the Boltz-2 affinity module: in the i.i.d. setting its RMSLE is lower by 0.433 log units, and it retains lower error than Boltz-2 in each OOD scenario. The same overall trend is observed for Spearman's rank correlation (Tables J2, J3), indicating that the improvement is not restricted to absolute EC_{50} error but also extends to the relative ordering of active pairs by potency.

The same-backbone comparisons between ASMI-Reg, ASMI-Cen and ASMI-DR help separate the effect of the learning formulation from that of the architecture. Direct EC_{50} regression with ASMI-Reg performs markedly worse than ASMI-DR, showing that the gain cannot be explained by the ASMI backbone alone. Censored regression with ASMI-Cen substantially improves over ordinary regression in most settings, confirming that explicitly accounting for inactive and censored observations is important. Nevertheless, ASMI-DR further improves over ASMI-Cen in every evaluated setting, reducing RMSLE by 0.205 log units in the i.i.d. case and by as much as 0.722 log units in *Molecule-Single*. These results isolate an additional empirical benefit from learning the concentration-conditional response beyond the improvement obtained from censoring alone.

6.3 Robustness to activity threshold, post-hoc curve fitting, and concentration processing

A potential concern is that EC_{50} error may depend on the activity threshold and recall since the evaluation is done over the intersection of true positive pairs. We therefore sweep the ASMI-DR activity threshold $\tau \in [0.02, 0.8]$ and examine the resulting recall-RMSLE trade-off. As can be seen in Figure 4, lower thresholds increase recall, as expected, while introducing less certain active pairs and generally increasing EC_{50} error. Nevertheless, ASMI-DR continues to outperform the DTA baselines in higher-recall regimes. For example, at threshold $\tau = 0.02$, ASMI-DR reaches recall comparable to DeepDTAGen while retaining an RMSLE advantage between 0.35 and 0.66 log units across the evaluated settings. DTIAM and ProSmith reach higher recall at this threshold, but with substantially larger EC_{50} errors. Together with the individual-true-positive evaluation in Section J.2,

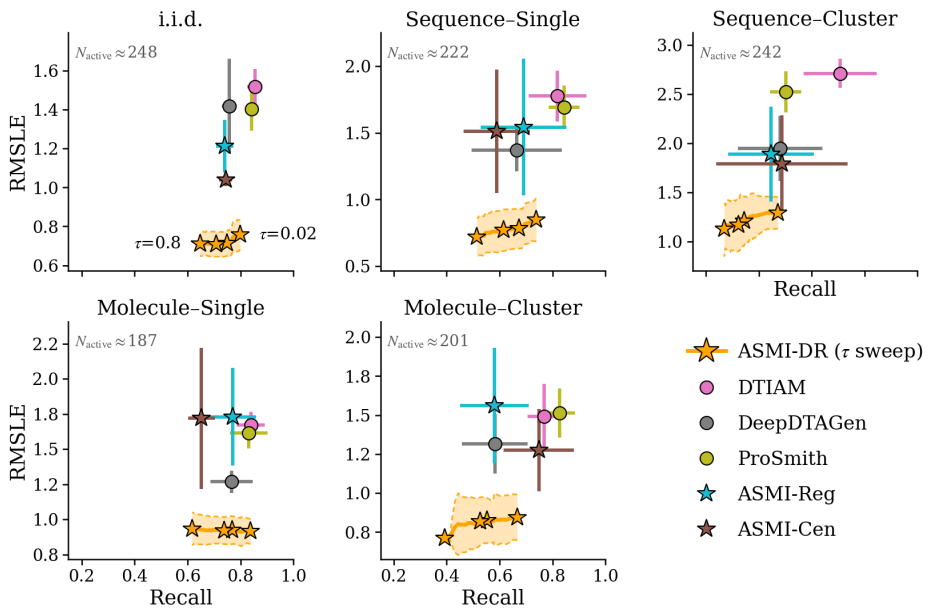


Figure 4: Recall-RMSLE trade-off across the activity decision threshold τ. ASMI-DR is swept over $\tau \in [0.02, 0.8]$ (orange). Recall decreases as τ increases, with the endpoints labelled in the i.i.d. panel. Stars mark $\tau \in \{0.02, 0.4, 0.5, 0.8\}$ and the shaded band is one standard deviation of RMSLE across the five splits. Each model is evaluated on its own true-positive activity predictions. N_{active} is the mean number of active test pairs per split, and error bars denote one standard deviation across splits. Lower RMSLE at equal recall is better. ASMI-DR does not reach the recall of DTIAM or ProSmith at any evaluated threshold, so the comparison in that range is an extrapolation.

this indicates that the EC_{50} gains are not solely a consequence of selecting a more conservative subset of active predictions.

We next test whether the improvement could be an artefact of the post-hoc procedure used to extract EC_{50} from the raw per-concentration predictions. Our default inference follows experimental practice by fitting a Hill curve and taking its inflection point. We additionally evaluate three alternatives: threshold crossing at a response of 0.5, isotonic regression, and a monotone I-spline fit. Across the i.i.d. and OOD settings, the resulting RMSLE values differ by less than approximately 0.1 log unit (Table F7). Hill fitting is generally the best-performing estimator, but the non-parametric alternatives yield closely comparable results. This shows that the improvement is robust to the specific procedure used to convert concentration-conditioned predictions into a scalar EC_{50} estimate.

Finally, we assess whether performance depends on the specific mechanism used to incorporate concentration into ASMI-DR. In a simpler late-concatenation variant, concentration is appended only before the output MLP rather than injected into the cross-attention layers. As shown in Table F5, this variant achieves RMSLE 0.724, compared with 0.676 for concentration-aware cross-attention, while attaining slightly higher MCC and recall. Thus, incorporating concentration directly into the interaction layers further improves EC_{50} estimation, but strong performance is retained even with a substantially simpler conditioning mechanism.

Together, these ablations indicate that the performance gains are mainly driven by the reformulation, and main improvement is robust to the activity decision threshold, the post-hoc EC_{50} extraction procedure, and the precise architectural implementation of concentration processing.

6.4 Activity prediction remains competitive

Although ASMI-DR is designed to recover potency from concentration-conditioned responses, it also provides an activity decision from the fitted response curve. We classify a pair as active when the fitted upper asymptote exceeds 0.5 and report MCC as the primary metric due to class imbalance.

Table 1: K_d estimation on five i.i.d. splits. RMSLE is calculated on the intersection of true positive activity predictions of ASMI-DR and DTA baselines, where for the regression baselines, we consider pairs with the estimated $K_d > 7.94\mu$M as inactive. Standard deviation is given in parentheses.

	Model	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
DAVIS	DTIAM	0.566 (0.014)	0.543 (0.013)	0.939 (0.008)	0.734 (0.018)	0.675 (0.010)
	DeepDTAGen	<u>0.642</u> (0.014)	0.691 (0.015)	0.816 (0.014)	0.796 (0.018)	0.622 (0.025)
	ProSmith	0.631 (0.010)	0.619 (0.013)	0.911 (0.007)	**0.681** (0.010)	**0.710** (0.008)
	ASMI-Reg	0.601 (0.048)	0.646 (0.063)	0.818 (0.030)	0.777 (0.070)	0.642 (0.033)
	ASMI-Cen	**0.685** (0.019)	0.818 (0.016)	0.722 (0.029)	0.778 (0.041)	0.641 (0.023)
	ASMI-DR	**0.685** (0.005)	0.772 (0.016)	0.775 (0.017)	<u>0.709</u> (0.019)	<u>0.694</u> (0.012)
BindingDB	DTIAM	0.606 (0.026)	0.678 (0.021)	0.964 (0.005)	<u>0.805</u> (0.020)	<u>0.772</u> (0.009)
	DeepDTAGen	0.722 (0.014)	0.827 (0.018)	0.879 (0.009)	0.863 (0.022)	0.744 (0.011)
	ProSmith	0.607 (0.041)	0.678 (0.033)	0.966 (0.008)	**0.801** (0.030)	**0.777** (0.015)
	ASMI-Reg	0.446 (0.231)	0.620 (0.140)	0.946 (0.043)	0.928 (0.128)	0.684 (0.095)
	ASMI-Cen	<u>0.727</u> (0.030)	0.873 (0.032)	0.823 (0.023)	0.926 (0.104)	0.688 (0.075)
	ASMI-DR	**0.745** (0.016)	0.839 (0.018)	0.890 (0.007)	0.833 (0.036)	0.767 (0.016)

When the curve cannot be fitted, we consider the pair to be inactive. For the regression baselines, we use EC_{50} as a decision variable, and we set pairs with the estimated $EC_{50} > 10^{-1}$M as inactive.

As can be seen in Figure 3 and Tables J2 and J3, ASMI-DR remains competitive in all evaluated scenarios. It performs particularly well in the *Sequence-Single* and *Molecule-Single* settings, while some classification or regression baselines are stronger in the more challenging cluster-level splits. Thus, the substantial gain in EC_{50} estimation does not come at the cost of the activity classification performance.

Structure-based baselines, including docking and Boltz-2, perform considerably worse for functional activity prediction in most settings, which may be due to the fact that binding is not a sufficient condition for activation in GPCRs (Zhang et al., 2024). When primary and secondary screening outcomes are treated as predictors of subsequent dose-response activity, ASMI-DR also performs better in the i.i.d., *Sequence-Single*, and *Molecule-Single* settings, while screening remains stronger in *Sequence-Cluster*. In case new molecular scaffolds are tested on previously explored proteins (*Molecule–Cluster*), ASMI-DR falls behind secondary screening but outperforms primary screening by a relative gain of 67% in MCC.

6.5 Generalisation to binding affinity

To assess whether the concentration-conditional formulation extends beyond GPCR functional responses, we evaluate ASMI-DR on the DAVIS and BindingDB affinity benchmarks. In analogy with EC_{50} from GPCR functional assays, K_d in log units obtained via competitive binding assays can be modelled by the generalised logistic curve. Thus, the framework can be adapted to model $P(\text{binding} \mid s, m, c)$ and recover the dissociation constant K_d from the concentration-dependent predictions, analogously to EC_{50} estimation.

As shown in Table 1, ASMI-DR achieves the highest activity MCC on both datasets while remaining competitive in K_d estimation. On DAVIS, its RMSLE is second to ProSmith, and on BindingDB it is the third-best-performing model, while substantially outperforming its direct-regression counterpart across main metrics. The censored regression variant of the architecture achieves on-par activity decision performance to ASMI-DR but lags behind in the K_d estimation. Moreover, MCC of ASMI-DR on DAVIS also matches that of the best dedicated classification DTI model (Table J6). These results indicate that the benefit of concentration-conditioned learning is not specific to olfactory receptor EC_{50} prediction and can transfer to standard binding-affinity settings.

7 Analysis and Limitations

ASMI-DR does not impose a fixed Hill shape during training. Instead, the supervision only encodes the concentration ordering implied by EC_{50}, and the raw predicted responses can therefore be non-sigmoidal. The Hill model is introduced only at inference time to extract a scalar EC_{50}, consistent with the wet lab protocols. Because training targets are derived from EC_{50} rather than directly observed responses at each concentration, the model output should primarily be interpreted as a concentration-conditioned response score rather than a calibrated probability of experimental activation. We examine the calibration of these scores in Section E, which provides additional insight into their probabilistic interpretation, reports the fitted curve parameters distribution, and compares predictions with available experimental dose-response measurements.

The flexibility of the framework comes at an additional inference cost because EC_{50} estimation requires evaluating several concentrations and fitting a curve. However, the number of queries can be reduced substantially: using seven concentrations changes performance by less than 0.015 RMSLE (Section G). In this setting, EC_{50} inference is approximately $12\times$ slower than direct regression, with 67% overhead due to independently parallelisable curve fitting. When only an activity decision is required, a single high-concentration query is sufficient, making inference comparable to standard single-pass baselines.

The current framework also has several scope limitations. Training relies on surrogate concentration-dependent labels derived from reported EC_{50} values rather than directly observed response measurements, and therefore the predicted curves should not be interpreted as reconstructions of the full experimental dose-response curve. In addition, the present formulation targets agonist activity for single protein–molecule pairs. Extending it to inhibitory responses, inverse agonism, allosteric modulation, and competitive multi-ligand settings remains future work. Single-point screening measurements could also be incorporated naturally through concentration-conditioned supervision, although screening data contain additional assay-specific sources of noise that would require careful treatment.

8 Conclusion

In this study, we propose a unified learning framework inspired by *in vitro* dose-response assays to simultaneously estimate activity and potency (EC_{50}) of protein-molecule pairs. Rather than assigning arbitrary potency values to inactive pairs, the proposed framework converts measured, censored, and inactive observations into concentration-dependent supervision and learns a response model from which both activity and potency can be recovered.

On M2OR, this formulation substantially improves EC_{50} estimation over standard regression, censored regression, and DTA baselines while retaining competitive activity prediction under i.i.d. and OOD evaluation. The approach also transfers to DAVIS and BindingDB, where it remains competitive for K_d estimation and improves activity discrimination. Together with the robustness analyses across activity thresholds, EC_{50} estimators, and concentration-conditioning mechanisms, these results suggest that the main gain comes from the concentration-conditional task formulation rather than a particular architectural or inference choice.

9 AI use statement

In this work, we used generative AI tools for minor coding assistance and as an aid to enhance writing clarity. We have not used generative AI tools for research ideation or execution, gathering and analysing literature, nor drafting entire sections of the paper. Application in synthetic data generation and proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work by thoroughly reading the outputs and applying our own independent judgement. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

10 Reproducibility statement

The anonymised code with example config files to train the models and run experiments is attached with the submission. Data preprocessing steps are described in Section H, model details are given in Section D, and model and baseline hyperparameters are described in Section D.1 and Section I, respectively.

References

P. Alonso Campana, P. Prasse, and T. Scheffer. Predicting dose-response curves with deep neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1144–1154. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/alonso-campana24a.html.

J. Ba, J.R. Kiros, and G.E. Hinton. Layer normalization. *arXiv*, abs/1607.06450, 2016. URL https://arxiv.org/abs/1607.06450.

T. K. Bjarnadóttir, D. E. Gloriam, S. H. Hellstrand, H. Kristiansson, R. Fredriksson, and H. B. Schiöth. Comprehensive repertoire and phylogenetic analysis of the g protein-coupled receptors in human and mouse. *Genomics*, 88(3):263–273, 2006. ISSN 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2006.04.001. URL https://www.sciencedirect.com/science/article/pii/S0888754306001054.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

C. Brideau, B. Gunter, B. Pikounis, and A. Liaw. Improved statistical methods for hit selection in high-throughput screening. *SLAS Discovery*, 8(6):634–647, 2003. ISSN 2472-5552. doi: https://doi.org/10.1177/1087057103258285. URL https://www.sciencedirect.com/science/article/pii/S2472630322002345.

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014.

X. Cong, W. Ren, J. Pacalon, R. Xu, L. Xu, X. Li, C.A. de March, H. Matsunami, H. Yu, Y. Yu, and J. Golebiowski. Large-scale G protein-coupled olfactory receptor–ligand pairing. *ACS Central Science*, 8(3):379–387, 2022. doi: 10.1021/acscentsci.1c01495. URL https://doi.org/10.1021/acscentsci.1c01495.

G. Corso, L. Cavalleri, D. Beaini, P. Liò, and P. Veličković. Principal neighbourhood aggregation for graph nets, 2020. URL https://arxiv.org/abs/2004.05718.

M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M. Hocker, D.K. Treiber, and P.P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 2011.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009. doi: 10.18637/jss.v032.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v032i05.

C.A. de March, J. Topin, E. Bruguera, G. Novikov, K. Ikegami, H. Matsunami, and J. Golebiowski. Odorant receptor 7D4 activation dynamics. *Angewandte Chemie International Edition*, 57 (17):4554–4558, 2018. doi: https://doi.org/10.1002/anie.201713065. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201713065.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Yuanqi Du, Arian R. Jamasb, Jeff Guo, Tianfan Fu, Charles Harris, Yingheng Wang, Chenru Duan, Pietro Liò, Philippe Schwaller, and Tom L. Blundell. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6(6):589–604, Jun 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00843-5. URL https://doi.org/10.1038/s42256-024-00843-5.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, and F. Herrera. *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018. ISBN 9783319980744. URL https://books.google.fr/books?id=8Fp0DwAAQBAJ.

J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, and G.E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272, 2017.

Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 10 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1072. URL https://doi.org/10.1093/nar/gkv1072.

A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

R. Gupta, A. Mittal, V. Agrawal, S. Gupta, K. Gupta, R.R.Jain, P. Garg, S.K. Mohanty, R. Sogani, H.S. Chhabra, V. Gautam, T. Mishra, D. Sengupta, and G. Ahuja. Odorify: A conglomerate of artificial intelligence–driven prediction engines for olfactory decoding. *Journal of Biological Chemistry*, 297 (2):100956, 2021. ISSN 0021-9258. doi: https://doi.org/10.1016/j.jbc.2021.100956. URL https://www.sciencedirect.com/science/article/pii/S0021925821007560.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2024. URL http://github.com/google/flax.

Franziska M. Heydenreich, Maria Marti-Solano, Manbir Sandhu, Brian K. Kobilka, Michel Bouvier, and M. Madan Babu. Molecular determinants of ligand efficacy and potency in gpcr signaling. *Science*, 382(6677):eadh1859, 2023. doi: 10.1126/science.adh1859. URL https://www.science.org/doi/abs/10.1126/science.adh1859.

Matej Hladiš, Maxence Lalis, Sebastien Fiorucci, and Jérémie Topin. Matching receptor to odorant with protein language and graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=q9VherQJd8_.

K. Huang, C. Xiao, L.M. Glass, and J. Sun. MolTrans: Molecular interaction transformer for drug-target interaction prediction. *Bioinformatics*, 37(6):830–836, 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa880. URL https://doi.org/10.1093/bioinformatics/btaa880.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, 2021. URL https://arxiv.org/abs/2102.09548.

M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T.L. Madden. NCBI BLAST: A better web interface. *Nucleic Acids Research*, 36(suppl_2):W5–W9, 2008.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

S. Junek, E. Kludt, F. Wolf, and D. Schild. Olfactory coding with patterns of response latencies. *Neuron*, 67(5):872–884, 2010. ISSN 0896-6273. doi: 10.1016/j.neuron.2010.08.005. URL https://doi.org/10.1016/j.neuron.2010.08.005.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Miroslav Kosar, Roman C. Sarott, David A. Sykes, Alexander E. G. Viray, Rosa Maria Vitale, Nataša Tomašević, Xiaoting Li, Rudolf L. Z. Ganzoni, Bilal Kicin, Lisa Reichert, Kacper J. Patej, Uxía Gómez-Bouzó, Wolfgang Guba, Peter J. McCormick, Tian Hua, Christian W. Gruber, Dmitry B. Veprintsev, James A. Frank, Uwe Grether, and Erick M. Carreira. Flipping the gpcr switch: Structure-based development of selective cannabinoid receptor 2 inverse agonists. *ACS Central Science*, 10(5):956–968, May 2024. ISSN 2374-7943. doi: 10.1021/acscentsci.3c01461. URL https://doi.org/10.1021/acscentsci.3c01461.

J. Kowalewski and A. Ray. Predicting human olfactory perception from activities of odorant receptors. *iScience*, 23(8):101361, 2020. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2020.101361. URL https://www.sciencedirect.com/science/article/pii/S2589004220305484.

Alexander Kroll, Sahasra Ranjan, and Martin J. Lercher. A multimodal transformer network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships. *PLOS Computational Biology*, 20(5):1–23, 05 2024. doi: 10.1371/journal.pcbi.1012100. URL https://doi.org/10.1371/journal.pcbi.1012100.

M. Lalis, M. Hladiš, S. Abi Khalil, L. Briand, S. Fiorucci, and J. Topin. M2OR: a database of olfactory receptor–odorant pairs for understanding the molecular mechanisms of olfaction. *Nucleic Acids Research*, 52(D1):D1370–D1379, 10 2024a. ISSN 0305-1048. doi: 10.1093/nar/gkad886. URL https://doi.org/10.1093/nar/gkad886.

M. Lalis, M. Hladiš, S. Abi Khalil, C. Deroo, C. Marin, M. Bensafi, N. Baldovini, L. Briand, S. Fiorucci, and J. Topin. A status report on human odorant receptors and their allocated agonists. *Chemical Senses*, 49:bjae037, 10 2024b. ISSN 1464-3553. doi: 10.1093/chemse/bjae037. URL https://doi.org/10.1093/chemse/bjae037.

S. Lee, I. Depoortere, and H. Hatt. Therapeutic potential of ectopic olfactory and taste receptors. *Nature Reviews Drug Discovery*, 18(2):116–138, 2019. ISSN 1474-1784. doi: 10.1038/s41573-018-0002-3. URL https://doi.org/10.1038/s41573-018-0002-3.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Zhangli Lu, Guoqiang Song, Huimin Zhu, Chuqi Lei, Xinliang Sun, Kaili Wang, Libo Qin, Yafei Chen, Jing Tang, and Min Li. Dtiam: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nature Communications*, 16(1):2548, Mar 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-57828-0. URL https://doi.org/10.1038/s41467-025-57828-0.

J.D. Mainland, A. Keller, Y.R. Li, T. Zhou, C. Trimmer, L.L. Snyder, A.H. Moberly, K.A. Adipietro, W.L.L. Liu, H. Zhuang, S. Zhan, S.S. Lee, A. Lin, and H. Matsunami. The missense of smell: functional variability in the human odorant receptor repertoire. *Nature Neuroscience*, 17(1):114–120, 2014. ISSN 1546-1726. doi: 10.1038/nn.3598. URL https://doi.org/10.1038/nn.3598.

B. Malnic, J. Hirono, T. Sato, and L.B. Buck. Combinatorial receptor codes for odors. *Cell*, 96(5): 713–723, 1999.

N. Malo, J.A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, 24(2):167–175, 2006. ISSN 1546-1696. doi: 10.1038/nbt1186. URL https://doi.org/10.1038/nbt1186.

Lily Masters, Scott Eagon, and Michael Heying. Evaluation of consensus scoring methods for autodock vina, smina and idock. *Journal of Molecular Graphics and Modelling*, 96:107532, 2020. ISSN 1093-3263. doi: https://doi.org/10.1016/j.jmgm.2020.107532. URL https://www.sciencedirect.com/science/article/pii/S1093326319307272.

Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 33–42, 2017. doi: 10.1109/ICDMW.2017.12.

K. Nara, L.R. Saraiva, X. Ye, and L.B. Buck. A large-scale analysis of odor coding in the olfactory epithelium. *Journal of Neuroscience*, 31(25):9179–9191, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1282-11.2011. URL https://www.jneurosci.org/content/31/25/9179.

R.R. Neubig, M. Spedding, T. Kenakin, and A. Christopoulos. International union of pharmacology committee on receptor nomenclature and drug classification. xxxviii. update on terms and symbols in quantitative pharmacology. *Pharmacological Reviews*, 55(4):597–606, 2003. ISSN 0031-6997. doi: 10.1124/pr.55.4.4. URL https://pharmrev.aspetjournals.org/content/55/4/597.

Y. Niimura and M. Nei. Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences*, 100(21):12235–12240, 2003. doi: 10.1073/pnas.1635157100. URL https://www.pnas.org/doi/abs/10.1073/pnas.1635157100.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707.

Jiao Qin, Ye Cai, Zheng Xu, Qianqian Ming, Su-Yu Ji, Chao Wu, Huibing Zhang, Chunyou Mao, Dan-Dan Shen, Kunio Hirata, Yanbin Ma, Wei Yan, Yan Zhang, and Zhenhua Shao. Molecular mechanism of agonism and inverse agonism in ghrelin receptor. *Nature Communications*, 13(1): 300, Jan 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-27975-9. URL https://doi.org/10.1038/s41467-022-27975-9.

Rodrigo Quiroga and Marcos A. Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLOS ONE*, 11(5):1–18, 05 2016. doi: 10.1371/journal.pone.0155183. URL https://doi.org/10.1371/journal.pone.0155183.

J. O. Ramsay. Monotone Regression Splines in Action. *Statistical Science*, 3(4):425 – 441, 1988. doi: 10.1214/ss/1177012761. URL https://doi.org/10.1214/ss/1177012761.

H. Saito, M. Kubota, R.W. Roberts, Q. Chi, and H. Matsunami. RTP family members induce functional expression of mammalian odorant receptors. *Cell*, 119(5):679–691, 2004.

H. Saito, Q. Chi, H. Zhuang, H. Matsunami, and J.D. Mainland. Odor coding by a mammalian receptor repertoire. *Science Signaling*, 2(60):ra9, 2009.

Pir Masoom Shah, Huimin Zhu, Zhangli Lu, Kaili Wang, Jing Tang, and Min Li. Deepdtagen: a multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. *Nature Communications*, 16(1):5021, May 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-59917-6. URL https://doi.org/10.1038/s41467-025-59917-6.

H. Spors and A. Grinvald. Spatio-temporal dynamics of odor representations in the mammalian olfactory bulb. *Neuron*, 34(2):301–315, 2002. ISSN 0896-6273. doi: 10.1016/S0896-6273(02)00644-X. URL https://doi.org/10.1016/S0896-6273(02)00644-X.

K. Sriram and P.A. Insel. G protein-coupled receptors as targets for approved drugs: How many targets and how many drugs? *Molecular Pharmacology*, 93(4):251–258, 2018. ISSN 0026-895X. doi: 10.1124/mol.117.111062. URL https://molpharm.aspetjournals.org/content/93/4/251.

J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1907382.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

J. Vig, A. Madani, L.R. Varshney, C. Xiong, R. Socher, and N.F. Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YWtLZvLmud7.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, Aug 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

C.D. Wilson, G.O. Serrano, A.A. Koulakov, and D. Rinberg. A primacy code for odor identity. *Nature Communications*, 8:1477, 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01432-4. URL https://doi.org/10.1038/s41467-017-01432-4.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237. URL https://doi.org/10.1021/acs.jcim.9b00237.

Xinke Zhan, Tiantao Liu, Changqing Yu, Yu-An Huang, Zhuhong You, and Shirley W. I. Siu. Maardti: a multi-perspective attention aggregation model for the prediction of drug–target interactions. *Digital Discovery*, pp. –, 2025. doi: 10.1039/D5DD00311C. URL http://dx.doi.org/10.1039/D5DD00311C.

Mingyang Zhang, Ting Chen, Xun Lu, Xiaobing Lan, Ziqiang Chen, and Shaoyong Lu. G protein-coupled receptors (gpcrs): advances in structures, mechanisms and drug discovery. *Signal Transduction and Targeted Therapy*, 9(1):88, Apr 2024. ISSN 2059-3635. doi: 10.1038/s41392-024-01803-6. URL https://doi.org/10.1038/s41392-024-01803-6.

Q. Zhao, H. Zhao, K. Zheng, and J. Wang. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3):655–662, 2021.

D. Zwicker. Primacy coding facilitates effective odor discrimination when receptor sensitivities are tuned. *PLOS Computational Biology*, 15(7):e1007188, 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007188. URL http://dx.doi.org/10.1371/journal.pcbi.1007188.

A TRAINING AND INFERENCE PSEUDO-CODE

Algorithm 1: Training

Input : Dataset $\mathcal{D} = \{(s^i, m^i), A^i, l^i\}_i$ with molecules m^i, sequences s^i, EC_{50} value A^i, and activity label l^i
Input : Concentration range (L, U)
Input : Margins $\varepsilon_L, \varepsilon_U$
X = [], Y = []
for $batch \sim \mathcal{D}$ **do**
  for i, (s, m), A, $label$, ***in*** *enumerate(batch)* **do**
    /* Set negative examples and truncate to (L,U)
    if $label = 0$ **then** $A \leftarrow U + \varepsilon_L$
    if $A > U + \varepsilon_L$ **then** $A \leftarrow U + \varepsilon_L$
    if $A < L - \varepsilon_U$ **then** $A \leftarrow L - \varepsilon_U$
    /* Sample a concentration c and a training label l_{tr}
    $c \sim Unif(L, U)$
    if $c \leq A - \varepsilon_L$ **then** $l_{tr} \leftarrow 0$
    else if $c \geq A + \varepsilon_U$ **then** $l_{tr} \leftarrow 1$
    else $l_{tr} \sim Unif(0, 1)$
    $X[i] \leftarrow (s, m, c)$
    $Y[i] \leftarrow l_{tr}$
  end
  /* Get batch sample weights based on probability of drawing a label for a given concentration.
  $W \leftarrow Get\_sample\_weights(batch)$
  /* Train as a standard classification task
  $\hat{Y} \leftarrow model(X)$
  $loss \leftarrow loss\_func(\hat{Y}, Y, W)$
  /* Take gradient of the loss and update model weights.
end

Algorithm 2: Inference

Input : Sequence-moleclue pair (s, m)
Input : Set C of concentrations uniformly covering range (L, U)
Function `DS_curve` (c, A, q, M, b) **:**
  return $(M - b)/(1 + pow(10, -q(c - A))) + b$
Y = []
for i, c ***in*** *enumerate(C)* **do**
  $Y[i] \leftarrow model((s, m, c))$
end
/* Use any standard algorithm to fit the curve
$\hat{A}, \hat{q}, \hat{M}, \hat{b} \leftarrow fit$(DS_curve, $x = C$, $y = Y$)
/* Return the estimated EC_{50} and the activation probability
return $\hat{A}$, $\hat{M}$

B SAMPLE WEIGHTS

Prediction of the protein-molecule response is inherently an imbalanced problem, and most of the experimental data are inactive. To account for the label distribution, we use standard imbalance ratio weights (Fernández et al., 2018). However, the sampling presented in Algorithm 1 dynamically generates labels in each iteration and changes the raw data label distribution, since active EC_{50} pairs can contribute to the inactive labels. Therefore, we calculate class weights on the fly for each batch by estimating the probability $P(l|c)$ that a label l is sampled for a given concentration c.

Formally, denote the truncated experimental EC_{50} values as $A^i_{trunc} = trunc(EC^i_{50}, (C_{low} - \epsilon_U, C_{high} + \epsilon_L))$. Then the probability that a label l is sampled for a concentration c in a batch of

size n is given by the number of times c is below or above the truncated EC_{50}:

$$P(l=0|c) \approx \frac{\#\{i|c \leq A^i_{trunc} - \epsilon_L\}}{n} \quad (2)$$

$$P(l=1|c) \approx \frac{\#\{i|c \geq A^i_{trunc} + \epsilon_U\}}{n} \quad (3)$$

$$P(l \in (0,1)|c) = 1 - P(l=0|c) - P(l=1|c) \quad (4)$$

Note that EC_{50} of inactive pairs is set to $+\infty$, thus $A^i_{trunc} = C_{high} + \epsilon_L$ for these pairs. Given the above estimate of the label probability, we set the class weights for each triplet (s, m, c) as

$$w_{class}(l, c) = \frac{1}{2P(l|c)} \quad (5)$$

C Training with EC_{50} lower bound

In the wet lab experiments, it is possible that an increase in the response has been observed in a dose-response assay, but the EC_{50} is outside the range of the tested concentrations. Thus, the curve cannot be fitted, and only a lower bound c_t of the EC_{50} is available. In such a case, our framework allows for a straightforward extension of the sampling procedure in Section 4 to this censored data.

Although EC_{50} value is not experimentally observed, the monotonicity of the dose-response curve implies that for all concentrations $c \leq c_t$ the response decision would be inactive. Therefore, "one-sided" training samples and their corresponding response decisions can be obtained for each protein-molecule pair i by uniformly sampling a concentration $c^{i,j} \in [C_{low}, C_{high}]$ and assigning $L^{i,j} = 0$ for $c^{i,j} \leq c_t$ and a soft label $L^{i,j} \sim Unif(0,1)$ for $c^{i,j} > c_t$. To further control the sampling, we introduce margins ϵ_L and ϵ_U and a summary of the sampling regions is given in Table C1.

Table C1: Sampling regions for all available data types and all possible training labels. ϵ_L, ϵ_U are lower and upper margins, respectively. The case $EC_{50} > c_t$ stands for EC_{50} experiments where activity was observed, but the EC_{50} is out of range of the tested concentrations, and the authors reported the lower bound c_t for the EC_{50} value.

Data type	Decision	$L^{i,j} = 0$	$L^{i,j} \sim Unif(0,1)$	$L^{i,j} = 1$
EC_{50}	active	$[C_{low}, EC_{50} - \epsilon_L]$	$(EC_{50} - \epsilon_L, EC_{50} + \epsilon_U)$	$[EC_{50} + \epsilon_U, C_{high}]$
EC_{50}	inactive	$[C_{low}, C_{high}]$	$\emptyset$	$\emptyset$
$EC_{50} > c_t$	active	$[C_{low}, c_t - \epsilon_L]$	$(c_t - \epsilon_L, C_{high}]$	$\emptyset$

D Model details

We represent the molecular input as a graph $m = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes (atoms) and $\mathcal{E}$ the set of edges (bonds). Each node and edge is initialised by feature vectors x_v and $e_{u,v}$, respectively, containing information such as atomic number, bond type, etc. (Table D1). To enhance the expressive power of the network, the graph is directed, and there are two edges $e_{u,v}$ and $e_{v,u}$ between each pair of nodes u and v (Yang et al., 2019).

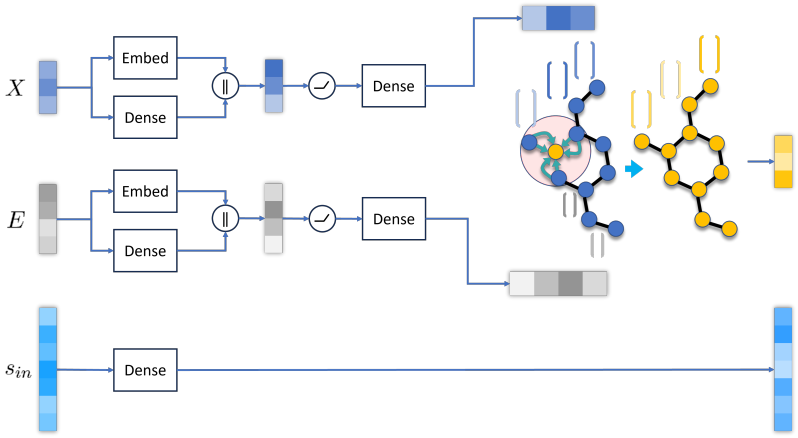


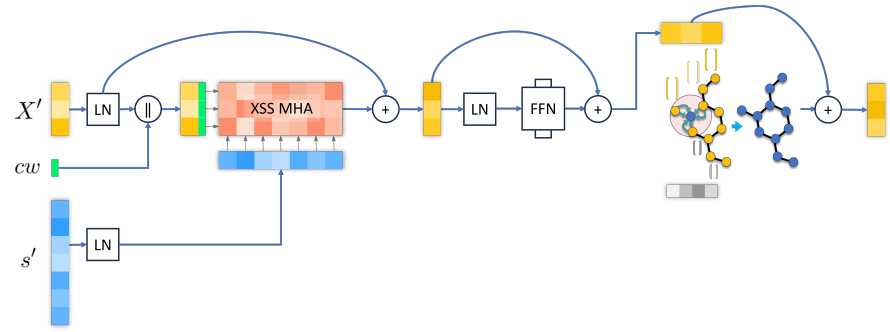
Figure D1: GNN embedding block. Node and edge feature matrices X and E, respectively, are embedded and processed by a message passing neural network (MPNN) (Gilmer et al., 2017). Categorical features are passed through a standard embedding layer, and continuous features are passed through a dense layer before being concatenated together. The sequence representation, comprising a d-dimensional vector for each amino acid, is passed through a single dense layer to form the input to the subsequent cross block. MPNN is composed of an edge update function that concatenates edge features, incoming node features, and outgoing node features and passes them through a dense layer followed by ReLU. It uses GRU (Cho et al., 2014) as a node update function. Symbols ⓘ, ⓒ stand for row-wise concatenation and the ReLU activation function, respectively.

Table D1: Initial node and edge features. *Cat.* stands for categorical features.

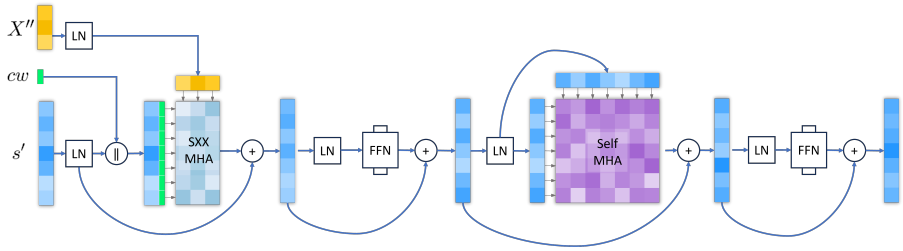
	Atom features	**Bond features**
Cat.	Atomic number Chiral tag Hybridisation Is aromatic	Bond type Stereo type Is aromatic
Continous	Formal charge Num. of implicit Hs Explicit valence Mass	

For protein sequence representation, we use ESM-2 (Lin et al., 2023) with frozen weights, and the input for a sequence of size n_s is the embedding of the last ESM-2 layer with dimensions $s_{in} \in \mathbb{R}^{n_s \times d}$, where d is the embedding size.

Combining molecule and protein inputs in an early stage of processing turns out to be beneficial for performance (Hladiš et al., 2023). We build upon this observation in the architecture outlined in Figure 2. The molecular graph is first transformed by the GNN embedding block (Figure D1) and processed together with the sequence embedding via a series of cross blocks. Each cross block is composed of a node update block (Figure D2a) and a sequence update block (Figure D2b). The core of these blocks is multi-head cross-attention, which learns changes in the molecular node and amino acid embeddings induced by the interactions with the sequence and molecule, respectively. In the node update block, node embeddings are used as queries, and amino acid embeddings are keys and values. The signal is then passed through the residual connection (He et al., 2015) and feed-forward network (FFN) and finally processed by Graph Isomorphism Network (GIN) (Xu et al., 2019). This allows the model to exploit the graph structure and edge information that are not considered in cross-attention.



(a) Node update block



(b) Sequence update block

Figure D2: Update blocks. (a) Node update block. The inputs are the node embedding matrix X', the sequence representation matrix s', and the concentration vector cw. X' is passed through LayerNorm (LN) (Ba et al., 2016) and concatenated with cw to become queries in multi-head cross-attention (XSS MHA) (Vaswani et al., 2017). Sequence representation s' serves as keys/values, and the output of cross-attention is added to the normalised node embeddings via a residual connection (He et al., 2015). The purpose of XSS MHA is to learn the difference in the node representation that is induced by the interaction with the protein. The resulting updated node embeddings are then processed by a feed-forward network (FFN) (Vaswani et al., 2017), and together with edge embeddings E form input to Graph isomorphism network (GIN) (Xu et al., 2019). (b) Sequence update block. Analogously, the inputs to the block are the node embedding matrix X'', the sequence representation matrix s', and the concentration vector cw. s' is passed through LayerNorm (LN) and concatenated with cw to become queries in multi-head cross-attention (SXX MHA). Node embeddings X'' are keys/values in cross-attention, and the output of SXX MHA is added to the normalised sequence representation. The goal of SXX MHA is to learn the difference in the sequence representation induced by the interaction with the molecule at the given concentration. The updated sequence embedding is then passed through FFN and then to a Transformer block.

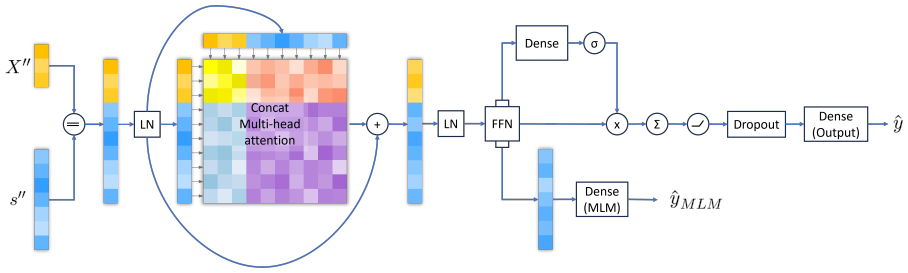


Figure D3: Final mixing block. The updated node embedding matrix X'' and the updated sequence representation s'' are concatenated together and form an input to self-attention (Vaswani et al., 2017). Note that unlike the case of update blocks, softmax in this self-attention is performed simultaneously through both molecular node and amino acid embeddings. The output of the attention layer is passed through FFN, aggregated via attention pooling (Eqn. (2) in (Hladiš et al., 2023)), and finally processed by a dense layer for a final activity prediction. The part of the FFN output corresponding to the sequence representation is also used for the masked language modelling task (MLM). The symbols ⊜, ⌣, σ, ×, Σ stand for column-wise concatenation, ReLU activation function, sigmoid activation function, element-wise multiplication and summation, respectively.

The sequence update block is designed analogously. The sequence embedding is transformed in the cross-attention layer, where amino acids are queries, and molecular nodes are keys and values. The updated sequence representation is then processed by a self-attention layer (Vaswani et al., 2017), which can be interpreted as GAT (Veličković et al., 2018) applied to a fully-connected graph where all amino acids are linked to each other.

An additional input to the cross blocks is concentration c, which is used in the form of auxiliary query features in cross-attention. It is first mapped to a learnt vector $cw \in \mathbb{R}^{d_c}$ for which the norm depends on the input concentration. Then cw is concatenated to each query just before multi-head cross-attention (green rectangle in Figure D2), allowing the model to learn the concentration-dependent protein-molecule interaction.

Finally, after N cross blocks, the embeddings are combined in the mixing block, followed by the output heads (Figure D3). Here, concatenated node and amino acid embeddings are passed to multi-head self-attention, followed by residual connection and FFN. In cross-attention, the softmax is either performed per node or per amino acid, and the model is restricted to learning the interaction between the molecule and the sequence. In contrast, softmax in the mixing block is performed through nodes and amino acids simultaneously and can give weight to self-interactions within the molecule and sequence (yellow and purple rectangles in Figure D3).

The final concatenated embedding from the mixing block is passed to a pooling layer[1] and MLP to obtain the final prediction. In addition, the sequence representation is extracted from the output of the mixing block and used for the masked language modelling (MLM) task (Figure D3) (Devlin et al., 2019). MLM has been shown to extract protein structure information from the sequence (Vig et al., 2021; Lin et al., 2023), and to avoid forgetting this information during training, we add the MLM task as an auxiliary training goal. We employ a common strategy to replace an amino acid by a [MASK] token with 15% chance. Apart from retaining the structural information, masking amino acids also serves as a training perturbation to avoid overfitting.

D.1 ASMI HYPERPARAMETERS

In the experiments, we use 2 cross blocks, sequences are padded to $n_s = 512$ and the sequence embedding is initialised by ESM-2 with 33 layers and $d = 1280$. Graph node embedding size was set to $d_{node} = 256$, edge embedding size to $d_{edge} = 32$ and concentration embedding size to $d_{conc} = 64$. We use 4 heads in all attention layers and a widening factor of 2 in the FFN layers. We use Trust region reflective algorithm implemented in SciPy v1.13.1 (Virtanen et al., 2020) to fit the curve (1) to the raw ASMI-DR predictions.

[1]We use attention pooling from Eqn. (2) in (Hladiš et al., 2023).

During the ASMI architecture development, we experimented with several choices, such as MPNN (Gilmer et al., 2017), GIN (Xu et al., 2019) or GAT (Veličković et al., 2018) as the choice of the GNN layer in Node update block, and PNA (Corso et al., 2020) and attention pooling (Hladiš et al., 2023) as the final pooling function. We did not run an exhaustive hyperparameter search.

D.2 ASMI-DR TRAINING

ASMI-DR is trained according to Algorithm 1 with $1,500$ epochs, a batch size of $1,024$, and margins $\epsilon_L = \epsilon_U = 0.25$. During training, we apply dropout of 0.1 in the attention layers, 0.2 in the FFN layers, and 0.5 just before the output layer. Padding is set to 32 nodes and 64 edges, which covers all molecules in the M2OR dataset. We use Adam Kingma & Ba (2015) optimizer and inverse square root decay scheduler with warmup Vaswani et al. (2017), the initial learning rate is set to $\frac{1}{\sqrt{256}}$, and we use $6,000$ warm-up steps for the scheduler. The best epoch was chosen based on the validation set comprising 10% of randomly selected pairs. All models are implemented in JAX (Bradbury et al., 2018) and FLAX, (Heek et al., 2024) and they were trained on Nvidia A100 SXM4 80GB GPUs or Nvidia H100 NVL 94GB GPUs.

D.3 ASMI-PROB TRAINING

To train the concentration-free version of the proposed architecture, we mostly follow Hladiš et al. (Hladiš et al., 2023). However, we drop Pair imbalance weights (Eqn. (5) in (Hladiš et al., 2023)) as these weights are related to the *in vitro* exploration of protein-molecule pairs, and we observe that these weights lower the performance. We also experimented with weights based on the receptor broadness (Lalis et al., 2024b) (data not shown), but their gain was not significantly different compared to standard class imbalance weights. We train ASMI-Prob for $1,500$ epochs with the batch size $1,024$, graph padding of 32 and 64 for nodes and edges, Adam Kingma & Ba (2015) optimizer, inverse square root decay scheduler with warmup Vaswani et al. (2017), the initial learning rate $\frac{1}{\sqrt{256}}$, and $6,000$ warm-up steps. The best epoch was chosen based on the validation set comprising 10% of randomly selected pairs.

D.4 ASMI-REG TRAINING

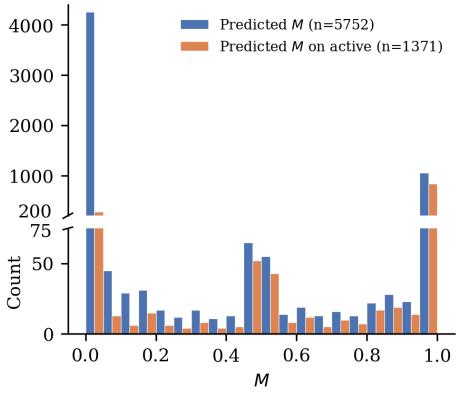
We train the regression variant of the proposed architecture similarly to ASMI-DR. We train for $1,500$ epochs with the batch size $1,024$, Adam Kingma & Ba (2015) optimizer, inverse square root decay scheduler with warmup Vaswani et al. (2017), the initial learning rate $\frac{1}{\sqrt{256}}$, $6,000$ warm-up steps, and graph padding of 32 and 64 for nodes and edges, respectively. We use l_2 loss to train the model. For inactive pairs, we set training EC_{50} values to 1M. As before, the best epoch was chosen based on the validation set comprising 10% of randomly selected pairs.

D.5 ASMI-CEN TRAINING

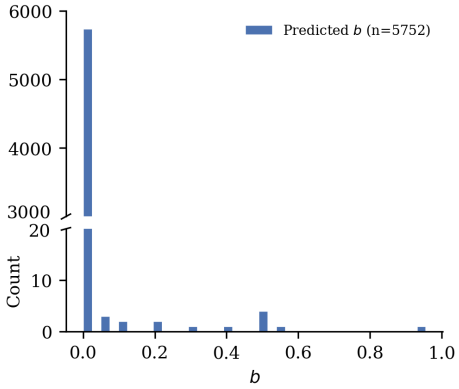
We train the censored regression variant of the proposed architecture with a right-censored type I Tobit likelihood (Tobin, 1958). We train for $1,500$ epochs with the batch size $1,024$, Adam Kingma & Ba (2015) optimizer, inverse square root decay scheduler with warmup Vaswani et al. (2017), the initial learning rate $\frac{1}{\sqrt{256}}$, $6,000$ warm-up steps, and graph padding of 32 and 64 for nodes and edges, respectively. To stabilise early training, the censored terms were excluded from the loss for the first 150 optimisation steps so that the model was first fitted to the measured EC_{50} values alone, after which training continued on the full Tobit likelihood. As before, the best epoch was chosen based on the validation set comprising 10% of randomly selected pairs.

E ANALYSIS OF THE FITTED CURVES

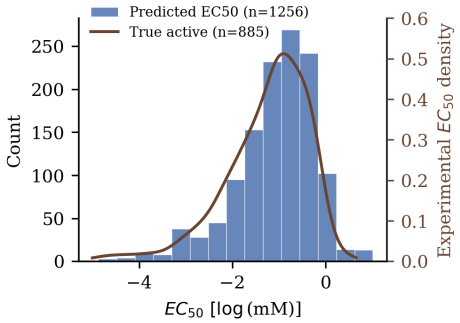
E.1 DISTRIBUTION OF FITTED HILL CURVE PARAMETERS



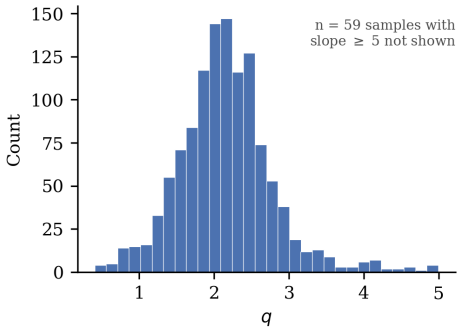
(a) Distribution of the upper asymptote M for all pairs (blue) and for experimentally active pairs (orange).



(b) Distribution of the lower asymptote b.



(c) EC_{50} distribution. The brown line corresponds to the experimentally obtained EC_{50} distribution.



(d) Distribution of slope q.

Figure E1: Distributions of the fitted Hill curve parameters derived from ASMI-DR predictions. The parameter distributions correspond to the test set pairs pooled across five i.i.d. splits.

E.2 CALIBRATION OF THE PREDICTED DOSE-RESPONSE CURVES

Although ASMI-DR is trained using surrogate concentration-dependent labels rather than raw experimental dose-response measurements, its output is intended to represent the progression from an inactive to an active response as concentration increases. We therefore assess whether the predicted response scores are calibrated with respect to experimentally observed dose-response measurements.

For active protein-molecule pairs for which raw dose-response data are available from Mainland et al. (2014), we compare the predicted $P(\text{active}|s, m, c)$ with the corresponding experimentally observed responses. Raw readouts for this analysis are min-max normalised, and following Mainland et al. (2014), the slope in (1) is fixed to $q = 1$. Margins in ASMI-DR training were set to $\epsilon_U = \epsilon_L = \log(\frac{0.9}{1-0.9})$, which corresponds to reaching 90% response at concentration $c = EC_{50} + \epsilon_u$ for slope $q = 1$ (i.e., $\frac{g(EC50+\epsilon_u)-b}{M-b} = 0.9$). All analysed pairs are excluded from the training sets, and the test set performance does not substantially change with this choice of the margins (MCC 0.666; RMSLE 0.753 log units).

Figure E2 shows that the predicted responses broadly follow the empirical dose-response magnitude across the probability range. Predictions in the intermediate range are close to the diagonal, whereas

larger deviations are observed at the extremes, most notably for high predicted responses. This behaviour can also be well observed in a set of representative examples in Figure E3. The resulting Brier score is 0.08. For reference, a dose-response curve fitted directly to the experimental measurements from Mainland et al. (2014) achieves a Brier score of 0.01, representing a substantially easier setting because the fit has direct access to the observations being evaluated. Thus, while ASMI-DR should not be interpreted as recovering the exact experimental response amplitude, its concentration-dependent output captures a meaningful approximation of the normalised dose-response progression despite being trained only from EC_{50}-derived supervision.

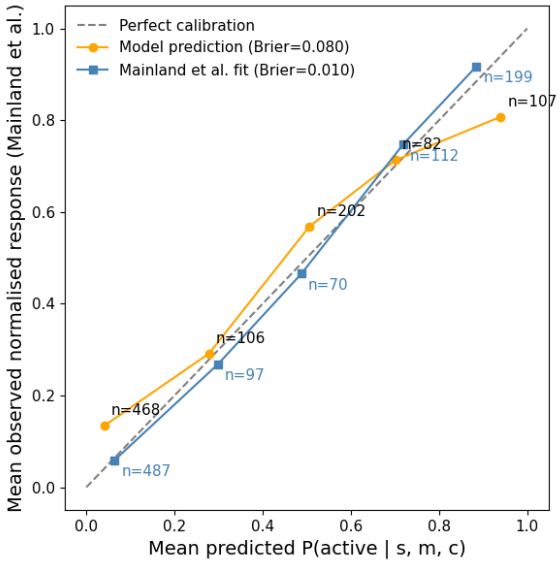


Figure E2: Calibration of concentration-conditional predictions against experimentally observed dose-response measurements. Mean ASMI-DR predictions $P(\text{active} \mid s, m, c)$ are compared with the corresponding mean normalised experimental responses from Mainland et al. (2014). The dashed line denotes perfect calibration. For reference, we also report calibration of a dose-response curve fitted directly to the experimental measurements (blue). Labels indicate the number of observations contributing to each point.

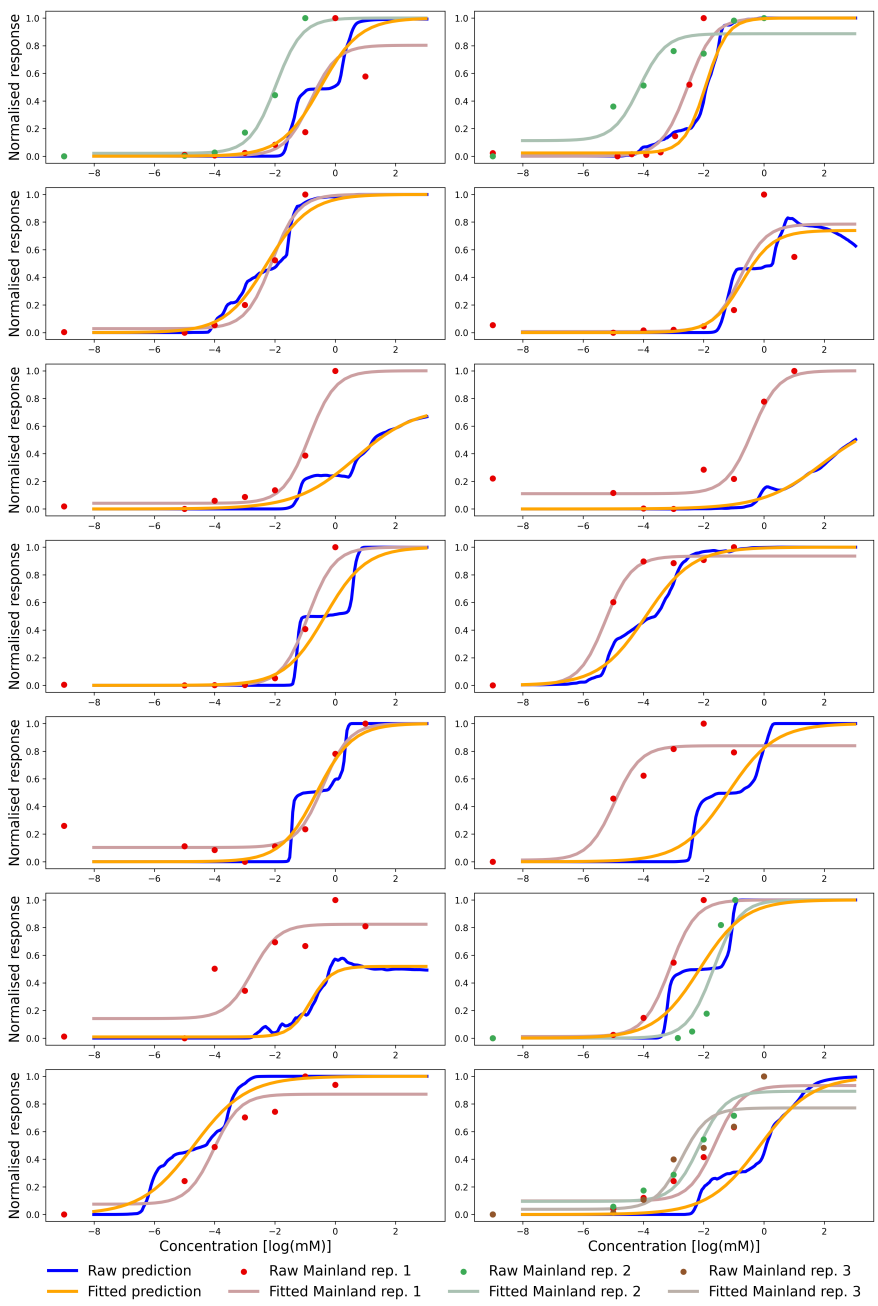


Figure E3: Examples of predicted curves corresponding to the correctly predicted active pairs for which the raw *in vitro* data are available. *rep.* stands for repetition.

F Ablation Studies

In the following sections, we evaluate the sensitivity of the training and inference pipelines on the choice of sampling margins ϵ_L and ϵ_U, soft label distribution in the region around EC_{50}, concentration sampling distribution, sample weighting strategy, concentration processing in the architecture, protein sequence featureizer, EC_{50} estimator, and activity decision threshold. Unless stated otherwise, all experiments are on i.i.d. setting of the M2OR dataset.

Overall, the results together suggest that the performance driver of ASMI-DR is concentration-conditional task reformulation rather than a particular choice of hyperparameters. While the perfor-

mance can change depending on the particular choices, all evaluated variants still outperform DTA baselines in EC_{50} estimation by a margin.

F.1 Ablations of training hyperparameters

Sensitivity analysis to the choice of ϵ. The sampling procedure in Algorithm 1 introduces hyperparameters ϵ_L and ϵ_U. To assess the robustness of our pipeline to the choice of these margins, we evaluate the performance when trained with $\epsilon_L = \epsilon_U = \epsilon \in \{0.1, 0.25, 0.5, 1.0\}$ and we report the results in Table F1. Overall, we found that ASMI-DR is robust to the choice of ϵ. The margins ϵ_L and ϵ_U defines the width of the transition window during sampling. When fitting the Hill model, varying this width primarily influences the slope q of the curve, rather than the inflection point. Consequently, the EC_{50} estimation remains stable. In the i.i.d. setting, MCC and RMSLE are consistent for $\epsilon \in \{0.1, 0.25, 0.5\}$ and MCC decreases only slightly at $\epsilon = 1.0$ (0.652), while RMSLE fluctuates around 0.729 ± 0.017 without a clear trend. Note that the reported RMSLE is evaluated on the true positive activity predictions of ASMI-DR only.

Table F1: Sensitivity analysis of the performance to ϵ in i.i.d. case. RMSLE is evaluated independently on true positives of each tested margin, and standard deviation is given in the parentheses.

ϵ	**MCC** ↑	**RMSLE** ↓
0.1	0.670 (0.013)	0.755 (0.093)
0.25	0.665 (0.011)	0.708 (0.064)
0.5	0.673 (0.027)	0.724 (0.074)
1.0	0.652 (0.034)	0.727 (0.050)

Soft label distribution in the transition region. Margins ϵ_L and ϵ_U introduce a transition region around EC_{50} and in Algorithm 1, we sample soft activity labels from uniform distribution $Unif(0, 1)$ for the concentrations in this region. Alternatively, one could replace this choice by drawing labels from an increasing line connecting the borders of the transition region $l = \frac{c-(EC_{50}-\epsilon_L)}{(EC_{50}+\epsilon_U)-(EC_{50}-\epsilon_L)}$ and we evaluate this option in Table F2.

Table F2: Ablation on the choice of soft label distribution in the transition region $c \in [EC_{50} - \epsilon_L, EC_{50} + \epsilon_U]$. EC_{50} estimation performance is calculated on the intersection of true positive pairs of the two approaches, and standard deviation is given in the parentheses.

Soft label	**MCC** ↑	**Precision** ↑	**Recall** ↑	**RMSLE** ↓	**Spearman's** ρ ↑
Linear	0.683 (0.023)	0.750 (0.031)	0.770 (0.009)	0.736 (0.112)	0.645 (0.063)
$Unif(0, 1)$	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.689 (0.063)	0.665 (0.049)

Upsampling concentration near EC_{50}. In our current design, concentration in Algorithm 1 is sampled uniformly from the experimental region $[C_{min}, C_{max}]$. However, the information content is not uniform, and activity labels around the boundaries give less information than the labels around EC_{50} where the transition occurs. Thus, we evaluate in Table F3 a variant of Algorithm 1 where for inactive pairs the sampling is unchanged and drawn from the uniform distribution and for active pairs the concentration is sampled uniformly with 50% probability and with 50% probability from a truncated normal distribution with $\mu = EC_{50}$ and $\sigma = 1$ and truncation interval $[C_{min}, C_{max}]$.

The results suggest that upsampling does not have any significant benefit, and two methods achieve comparable performance in both activity decision and EC_{50} estimation.

Table F3: Ablation on the concentration sampling during training. EC_{50} estimation performance is calculated on the intersection of true positive pairs of the two approaches, and standard deviation is given in the parentheses.

Conc. sampling	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
Upsampling	0.662 (0.028)	0.761 (0.038)	0.721 (0.033)	0.687 (0.096)	0.661 (0.093)
Uniform	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.664 (0.079)	0.669 (0.073)

The choice of sample weights. The proposed concentration-conditional formulation allows responsive pairs to contribute to the inactive labels for low concentrations. This changes the label distribution, and we compensate for this change using dynamic class weights described in Section B. To evaluate the benefit of this procedure, we compare its performance to the performance of ASMI-DR trained with standard static imbalance ratio weights, and we report the results in Table F4.

Table F4: Ablation on the dynamic class weights. EC_{50} estimation performance is calculated on the intersection of true positive pairs of the two approaches, and standard deviation is given in the parentheses.

Class weights	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
Imbalance	0.665 (0.025)	0.719 (0.072)	0.782 (0.058)	0.742 (0.055)	0.641 (0.057)
Dynamic	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.703 (0.066)	0.663 (0.058)

F.2 Model ablations

Late concentration concatenation. Concentration c is incorporated into the ASMI-DR architecture as auxiliary query features in the cross-attention mechanism. To quantify the benefit of this design choice, we evaluate a simplified variant in which the concentration is concatenated with the representation produced by the mixing block and the resulting vector is passed through a two-layer MLP. As shown in Table F5, cross-attention variant improves EC_{50} estimation, reducing the RMSLE from 0.724 with late concatenation to 0.676, at the cost of a 0.049 decrease in recall. These results also highlight our main contribution of task reformulation since most of the performance improvement over DTA baselines stems from the concentration-conditional formulation and not from the treatment of the concentration in the model architecture.

Table F5: Ablation on the concentration processing in the ASMI-DR architecture. *Concat.* stands for late concentration concatenation and *Cross-attn.* is currently used concentration as auxiliary query features in cross-attention. EC_{50} estimation performance is calculated on the intersection of true positive pairs of the two approaches, and standard deviation is given in the parentheses.

Variant	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
Concat.	0.688 (0.013)	0.767 (0.022)	0.757 (0.032)	0.724 (0.066)	0.657 (0.070)
Cross-attn	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.676 (0.074)	0.676 (0.061)

Sequence featurizer. Protein representation in ASMI-DR relies on a pre-trained protein language model, and as an alternative to ESM-2, we report in Table F6 the performance of ASMI-DR using ProtBERT Elnaggar et al. (2021). As can be seen, ESM-2 achieves lower RMSLE at the expense of a marginally lower MCC.

Table F6: Comparison of ProtBERT and ESM-2 as protein featurizer in ASMI-DR. EC_{50} estimation performance is calculated on the intersection of true positive pairs of the two approaches, and standard deviation is given in the parentheses.

Featurizer	**MCC** ↑	**Precision** ↑	**Recall** ↑	**RMSLE** ↓	**Spearman's** ρ ↑
ProtBERT	0.670 (0.023)	0.735 (0.036)	0.766 (0.016)	0.760 (0.055)	0.647 (0.070)
ESM-2	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.660 (0.055)	0.681 (0.066)

F.3 Comparison of EC_{50} estimators

Fitting a Hill curve and retrieving its inflection point is a natural formulation for the EC_{50} estimation, which follows standard wet lab protocols. In the following section, we further probe the effect of the post-hoc curve-fitting on the EC_{50} estimation using 3 alternatives: threshold crossing (TC), isotonic regression (Iso), and monotone third-degree I-spline fit (I-spline):

- *Threshold crossing* is a non-parametric estimate that fits a piecewise-linear curve and retrieves EC_{50} as the first interpolated concentration at which the probability of activation reaches 50%.
- *Isotonic regression* is a non-parametric estimate that uses isotonic regression (de Leeuw et al., 2009) to enforce a non-decreasing dose-response curve and retrieves EC_{50} as the interpolated concentration at which the probability of activation reaches 50%.
- *Monotone I-spline* is a non-parametric smooth estimate that fits a monotone I-spline (Ramsay, 1988) to the predicted dose-response curve and retrieves EC_{50} as the concentration at which the fitted probability of activation reaches 50%.

The results are given in Table F7. Overall, Algorithm 2 is robust to the choice of the EC_{50} estimator, with differences spanning below 0.1 log units. Generally, the best-performing estimator is the Hill curve fit usually used to treat raw wet lab experimental data, followed by a simple threshold crossing.

Table F7: Comparison of EC_{50} estimation methods. RMSLE and Spearman's ρ are evaluated independently on true positives of each estimator, and standard deviation is given in the parentheses.

		Estimator	**MCC** ↑	**Precision** ↑	**Recall** ↑	**RMSLE** ↓	**Spearman's** ρ ↑
i.i.d.		TC	0.670 (0.010)	0.771 (0.024)	0.723 (0.025)	0.708 (0.096)	0.640 (0.070)
		Iso	0.669 (0.011)	0.779 (0.024)	0.711 (0.027)	0.715 (0.083)	0.636 (0.062)
		I-spline	0.686 (0.005)	0.782 (0.025)	0.737 (0.033)	0.734 (0.074)	0.634 (0.061)
		Hill curve	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.708 (0.064)	0.650 (0.061)
Sequence	**Single**	TC	0.484 (0.023)	0.637 (0.044)	0.627 (0.081)	0.806 (0.121)	0.471 (0.110)
		Iso	0.487 (0.026)	0.645 (0.045)	0.619 (0.080)	0.803 (0.124)	0.471 (0.110)
		I-spline	0.505 (0.021)	0.647 (0.046)	0.648 (0.054)	0.836 (0.103)	0.452 (0.101)
		Hill curve	0.481 (0.031)	0.642 (0.038)	0.614 (0.095)	0.769 (0.156)	0.470 (0.119)
	Cluster	TC	0.210 (0.044)	0.440 (0.154)	0.336 (0.122)	1.242 (0.274)	0.018 (0.184)
		Iso	0.223 (0.047)	0.469 (0.166)	0.323 (0.128)	1.220 (0.309)	0.051 (0.202)
		I-spline	0.225 (0.050)	0.470 (0.167)	0.325 (0.131)	1.232 (0.313)	0.024 (0.164)
		Hill curve	0.217 (0.045)	0.461 (0.162)	0.324 (0.131)	1.168 (0.268)	0.040 (0.150)
Molecule	**Single**	TC	0.597 (0.079)	0.654 (0.097)	0.759 (0.111)	0.957 (0.096)	0.461 (0.124)
		Iso	0.598 (0.074)	0.666 (0.096)	0.743 (0.110)	0.958 (0.103)	0.465 (0.128)
		I-spline	0.609 (0.071)	0.669 (0.099)	0.758 (0.096)	0.965 (0.094)	0.463 (0.110)
		Hill curve	0.593 (0.074)	0.663 (0.098)	0.738 (0.104)	0.920 (0.096)	0.474 (0.116)
	Cluster	TC	0.407 (0.080)	0.567 (0.125)	0.551 (0.131)	0.861 (0.189)	0.277 (0.121)
		Iso	0.399 (0.080)	0.573 (0.121)	0.528 (0.147)	0.878 (0.179)	0.294 (0.100)
		I-spline	0.400 (0.080)	0.572 (0.121)	0.530 (0.144)	0.888 (0.169)	0.292 (0.117)
		Hill curve	0.398 (0.076)	0.572 (0.114)	0.526 (0.149)	0.818 (0.154)	0.298 (0.115)

F.4 Decision threshold sweep and recall-vs-RMSLE evaluation

To evaluate the sensitivity of EC_{50} prediction to the choice of activity decision τ, we report the performance over a sweep of thresholds $\tau \in [0.02, 0.8]$ and compare the results to the recall-vs-RMSLE properties of DTA baselines.

As can be seen in Table F8 and Figure 4, RMSLE is either stable or increases with the lowering thresholds and increasing recall. This is expected as the EC_{50} estimates of less certain active pairs is becoming more noisy. Notably, when comparing the results to Table J4, ASMI-DR outperforms dedicated DTA baselines by a margin even for higher recall and low values of τ. For instance, DeepDTAGen achieves lower recall for all evaluated cases except *Sequence-Cluster* while suffering from higher EC_{50} estimation errors. The relative gains of ASMI-DR against DeepDTAGen in RMSLE range from 0.35 log units in the *Molecule-Single* case up to 0.66 log units in the i.i.d. case. DTIAM and ProSmith achieve higher recall and lower precision than ASMI-DR with the lowest evaluated decision threshold $\tau = 0.02$, but their EC_{50} is still substantially higher than the error of ASMI-DR.

Table F8: Evaluation of AMSI-DR across decision thresholds τ. RMSLE and Spearman's ρ are evaluated independently on true positives of each case, and standard deviation is given in the parentheses.

		Name	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
i.i.d.		$\tau = 0.02$	0.674 (0.009)	0.733 (0.025)	0.776 (0.026)	0.758 (0.076)	0.616 (0.068)
		$\tau = 0.40$	0.682 (0.012)	0.773 (0.019)	0.741 (0.036)	0.712 (0.065)	0.649 (0.062)
		$\tau = 0.50$	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.708 (0.064)	0.650 (0.061)
		$\tau = 0.80$	0.635 (0.013)	0.791 (0.027)	0.642 (0.034)	0.711 (0.062)	0.644 (0.066)
Sequence	**Single**	$\tau = 0.02$	0.486 (0.030)	0.580 (0.047)	0.732 (0.045)	0.847 (0.160)	0.404 (0.115)
		$\tau = 0.40$	0.516 (0.031)	0.647 (0.037)	0.670 (0.045)	0.781 (0.150)	0.455 (0.110)
		$\tau = 0.50$	0.481 (0.031)	0.642 (0.038)	0.614 (0.095)	0.769 (0.156)	0.470 (0.119)
		$\tau = 0.80$	0.459 (0.022)	0.682 (0.075)	0.519 (0.073)	0.718 (0.147)	0.505 (0.112)
	Cluster	$\tau = 0.02$	0.226 (0.044)	0.417 (0.128)	0.433 (0.192)	1.292 (0.162)	0.007 (0.112)
		$\tau = 0.40$	0.233 (0.059)	0.475 (0.158)	0.335 (0.153)	1.208 (0.251)	0.027 (0.144)
		$\tau = 0.50$	0.217 (0.045)	0.461 (0.162)	0.324 (0.131)	1.168 (0.268)	0.040 (0.150)
		$\tau = 0.80$	0.208 (0.039)	0.486 (0.157)	0.258 (0.097)	1.124 (0.280)	0.014 (0.199)
Molecule	**Single**	$\tau = 0.02$	0.531 (0.170)	0.578 (0.181)	0.811 (0.114)	0.913 (0.089)	0.467 (0.122)
		$\tau = 0.40$	0.600 (0.072)	0.655 (0.111)	0.766 (0.108)	0.922 (0.099)	0.474 (0.104)
		$\tau = 0.50$	0.593 (0.074)	0.663 (0.098)	0.738 (0.104)	0.920 (0.096)	0.474 (0.116)
		$\tau = 0.80$	0.548 (0.065)	0.688 (0.068)	0.617 (0.069)	0.932 (0.118)	0.464 (0.119)
	Cluster	$\tau = 0.02$	0.394 (0.146)	0.534 (0.152)	0.621 (0.170)	0.844 (0.147)	0.245 (0.149)
		$\tau = 0.40$	0.404 (0.081)	0.567 (0.121)	0.548 (0.155)	0.824 (0.145)	0.285 (0.141)
		$\tau = 0.50$	0.398 (0.076)	0.572 (0.114)	0.526 (0.149)	0.818 (0.154)	0.298 (0.115)
		$\tau = 0.80$	0.320 (0.109)	0.548 (0.114)	0.392 (0.146)	0.709 (0.013)	0.266 (0.151)

G Inference time estimation

ASMI-DR inference is more computationally demanding than direct regression when estimating EC_{50}, since it requires evaluating the model at multiple concentrations and fitting a Hill curve. However, the number of queried concentrations provides a direct trade-off between runtime and accuracy.

In the main experiments on M2OR dataset, we queried 62 concentrations over the range 10^{-8}M to 10^{-2}M with a step of 0.1 log units to obtain a fine-grained curve fit. We find that this can be reduced substantially: using only 7 concentrations preserves performance within 0.015 RMSLE of the full inference procedure (Figure G1). In a 7-point setup with batch size 128, inference takes 0.331s per batch on an H100 94GB GPU with post-hoc curve fitting on an Intel Core i9-13900H CPU. This consists of 0.109s for the seven model evaluations (7×0.0156s) and 0.221s for CPU curve fitting. By comparison, ASMI-Reg requires 0.027s for a single batch. Thus, high-fidelity ASMI-DR EC_{50} estimation is approximately $12\times$ slower than direct regression, but remains practical:

processing 10^6 protein–molecule pairs would take approximately 43 minutes, corresponding to $\sim$380 pairs/s. Moreover, 67% of the runtime is spent on curve fitting, which is independent across pairs and therefore readily parallelisable across CPUs.

For virtual screening, where only a binary activity decision is required, ASMI-DR does not require curve fitting. A single forward pass at a high concentration is sufficient, reducing runtime to 0.0156s per batch. This introduces only a small decrease in activity prediction performance, with an average MCC drop of approximately 0.008. In this setting, ASMI-DR has inference cost comparable to standard single-pass regression or classification baselines.

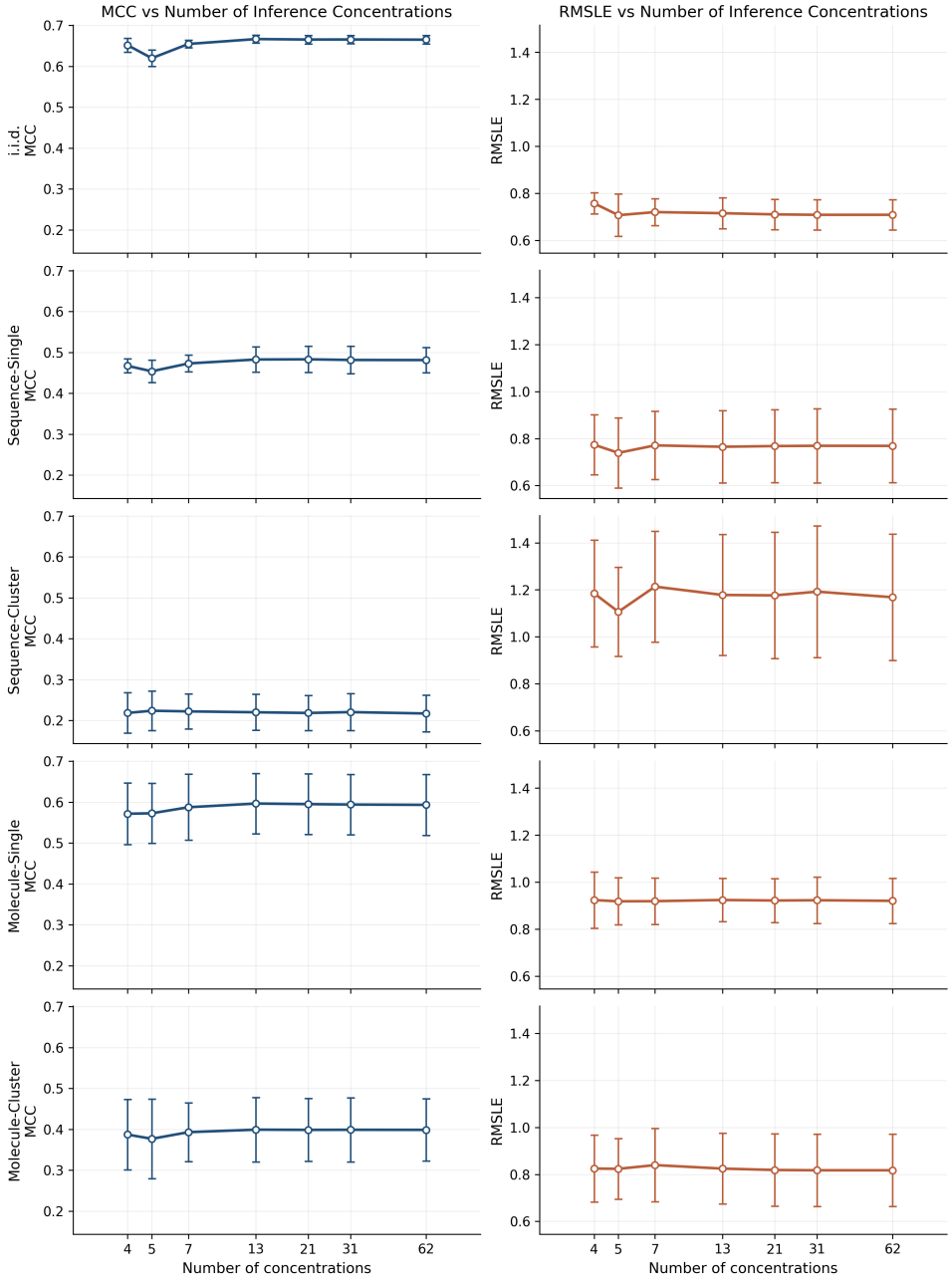


Figure G1: Number of inference concentrations vs the performance of ASMI-DR on M2OR.

H DATA PREPROCESSING

H.1 M2OR

M2OR gathers $77,611$ experiments, some of which correspond to the same protein-molecule pairs and $71,454$ are screening data. During preprocessing, we first discard mixtures of molecules and experiments measuring the basal activity (i.e., experiments performed at a molecular concentration of 0M). We also discard data reporting "fold" and "micro-ampere" units. We further remove the data with inconsistent activity decisions between experiments. For screening data, we remove experiments where, for a given protein-molecule pair, the activity decision in the higher concentration is "inactive" and in a lower concentration it is "active". We exclude galaxolide due to its unspecific response (Lalis et al., 2024b), and we also exclude molecules with molecular graphs composed of more than 32 nodes or 64 directed edges. In addition, we exclude putative pseudogenes with sequence lengths lower than 296 amino acids. Finally, we change the units to $log_{10}(mM)$, and if there are multiple dose-response experiments for a given pair, we take the mean of the EC_{50} values. After preprocessing, the dataset consists of $1,427$ EC_{50} values for active pairs, $4,346$ inactive dose-response pairs, and $60,256$ screening samples.

H.1.1 DATA SPLITTING DETAILS

In order to thoroughly test the performance, we adopt the evaluation procedure from Hladiš et al. Hladiš et al. (2023). Each scenario consists of 5 test sets, and each test set contains roughly 20% of the data. We describe the details of the five evaluation scenarios below:

- **i.i.d** corresponds to the standard random split evaluation procedure where each protein-molecule pair is held out in the test set with probability $1/N_{pairs}$.
- **Sequence–Single** evaluates the performance of the model when a new protein sequence is tested. The split is implemented by uniformly randomly selecting 20% of the protein sequences which have at least 2 known active compounds. Then all pairs corresponding to the selected sequences are held out in the test set.
- **Sequence–Cluster** assesses the generalisability to novel protein families. To construct the test sets, we use BLAST-derived sequence identity Johnson et al. (2008) and apply HDBSCAN McInnes & Healy (2017) to obtain clusters, which avoids forcing outliers into artificial groups. We retain candidate clusters with at least 2 sequences with at least 2 active compounds. We then randomly select 5 clusters, and all pairs corresponding to the sequences in these clusters are held out. The split is accepted if the test set size is between 12% and 40% of the total dataset and the procedure is repeated otherwise.
- **Molecule–Single** tests the model on previously unseen molecules. The split is constructed by uniformly randomly selecting 20% of the molecules which activate at least 2 protein sequences. Then all pairs corresponding to the selected molecules are held out in the test set.
- **Molecule–Cluster** assesses the generalisability to novel molecular scaffolds, and it is constructed analogously to the Sequence–Cluster scenario. We use the Tanimoto distance and apply HDBSCAN McInnes & Healy (2017) to obtain clusters of molecules. We then randomly select 5 clusters from candidate clusters with at least 2 molecules with at least 2 activated sequences. Then all pairs corresponding to the molecules in the selected clusters are held out in the test set. We accept the split if the test set size is between 12% and 40% of the total dataset and repeat the procedure otherwise.

We note that all splits and their hyperparameters and random seeds are provided in the released data and code for full reproducibility.

Activity label distribution of protein-molecule interaction datasets is inherently imbalanced, and we report the class ratio of all the evaluation scenarios in Table H1. As can be seen, the average ratio of active and inactive pairs remains stable across different split strategies. The standard deviation of Sequence–Single and Sequence–Cluster test sets is higher due to the imbalance in "broadness", i.e., in the number of active molecules per protein (Lalis et al., 2024b).

Table H1: Ratio of active/inactive data points in the evaluation scenarios of M2OR. Standard deviation is given in parenthesis.

	train set	valid set	test set
i.i.d	0.330 (0.007)	0.351 (0.034)	0.314 (0.026)
Sequnece–Single	0.315 (0.017)	0.321 (0.050)	0.411 (0.124)
Sequence–Cluster	0.353 (0.048)	0.371 (0.061)	0.347 (0.196)
Molecule–Single	0.336 (0.020)	0.332 (0.038)	0.318 (0.094)
Molecule–Cluster	0.331 (0.019)	0.325 (0.044)	0.325 (0.094)

H.2 DAVIS

We downloaded the curated data from the PyTDC library (Huang et al., 2021). For consistency with the previous work, we obtain the labels from the preprocessed data from the HyperAttentionDTI publication (Zhao et al., 2021). We exclude 6 protein sequences longer than $1,736$ amino acids, and we exclude molecular graphs with more than 128 nodes or 256 directed edges. Finally, we changed the units to $log_{10}(\mu M)$. After preprocessing, the data contains $6,881$ active and $17,667$ inactive pairs.

H.3 BINDINGDB

We downloaded the curated data from the PyTDC library (Huang et al., 2021). If there are duplicated experiments for the same pair, we take the lowest K_d. After analysing the data distribution, and in line with standard practices (Huang et al., 2021), we consider pairs with $K_d \geq 10\mu$M as inactive. We exclude 12 protein sequences longer than $2,048$ amino acids, and we exclude molecular graphs with more than 128 nodes or 256 directed edges. Finally, we changed the units to $log_{10}(\mu M)$. After preprocessing, the data contains $19,233$ active and $23,001$ inactive pairs.

H.4 KIBA

We downloaded the preprocessed dataset from the HyperAttentionDTI publication (Zhao et al., 2021). We exclude 2 protein sequences longer than $1,408$ amino acids, and we exclude molecular graphs with more than 128 nodes or 256 directed edges. After preprocessing, the data contains $22,154$ active and $94,195$ inactive pairs.

I BASELINE TRAINING AND EVALUATION DETAILS

I.1 *In vitro* EC_{50} ERROR

Measuring EC_{50} *in vitro* is prone to errors arising from several different sources (Malo et al., 2006; Brideau et al., 2003). In this work, this error is estimated by root mean squared logarithmic error (RMSLE) based on the protein-molecule pairs for which more than one dose-response experiment has been conducted. Formally, the error is given by the differences between individual EC_{50} values and the mean EC_{50} corresponding to the same protein-molecule pair:

$$\text{Experimental RMSLE} = \sqrt{\frac{1}{L}\sum_{k=1}^{K}\sum_{i=1}^{I_k}(x_{i,k} - \mu_k)^2} \tag{6}$$

where K is the number of protein-molecule pairs with multiple EC_{50} experiments, I_k is the number of experiments per pair k, $x_{i,k}$ is the individual EC_{50} value i for the pair k in logarithmic scale, $\mu_k = \sum_{i=1}^{I_k} x_{i,k}$ is the mean of the experiments for pair k, and $L = \sum_{k=1}^{K} I_k$ is the total number of experiments for protein-molecule pairs with multiple EC_{50} values in the dataset.

I.2 DOCKING

Gypsum-DL (v1.2.0) was used to generate 3D molecular structures from SMILES to SDF, accounting for ionisation, tautomeric, chiral, cis/trans, and ring-conformational states at pH 7.0 ± 0.5. Structures were converted from SDF to MOL2 with Open Babel 3.1.0 and then to PDBQT using MGLTools (v1.5.7). Docking was performed with SMINA (Oct 15, 2019, based on AutoDock Vina 1.1.2) (Masters et al., 2020) using the Vinardo scoring function (Quiroga & Villarreal, 2016) and an exhaustiveness of 8. Olfactory receptor models were obtained from the AlphaFold DB (Varadi et al., 2023; 2021). Polar hydrogens were added with PDB2PQR (v3.6.1), with protonation states assigned by PROPKA (v1.0) at pH 7.0, followed by minimisation with the AMBER99 force field. Protein structures were converted from PQR to PDBQT using MGLTools, and all receptors were superimposed with PyMOL (v2.5.4).

We perform docking on 868 (300 active and 568 inactive) pairs corresponding to the OR sequences available at AlphaFold DB. We generated up to nine poses and used the Youden index estimated on 10% of the data to transform docking scores into activity decisions.

I.3 BOLTZ-2

Olfactory receptor sequences and molecular SMILES were obtained from the M2OR database. We performed the experiments with Boltz-2 default parameters except for the following: the number of diffusion samples was set to 25 for both the structure and affinity modules, and the number of recycling steps was set to 10.

I.4 DTI AND DTA BASELINES

For all DTI and DTA baselines, we use the default hyperparameters reported in the corresponding papers for all the datasets. The only exception is MAARDTI, for which we use the hyperparameters reported by Zhan et al. (2025) for the KIBA dataset (Protein MAAR Dimension set to $c_p = 8$ and Drug MAAR Dimension set to $c_d = 16$).

J EXTENDED RESULTS

J.1 EXTENDED RESULTS ON M2OR DATASET

In this section, we report in Table J2 and Table J3 the underlying values presented visually in Figure 3, and to further interpret the results, we also report the number of active pairs and the size of the intersection of true positives across models and evaluation scenarios in Table J1.

Table J1: Average number of samples, active pairs, and true positive pairs in the intersection of ASMI-DR and DTA models.

	N	$\mathbf{N_{active}}$	$\frac{N_{active}}{N}$	**TP**	$\frac{TP}{N_active}$
i.i.d.	1127.2	248.2	0.220	153.2	0.617
Sequence-Single	848.2	222.4	0.261	118.0	0.517
Sequence-Cluster	1314.2	241.6	0.214	32.2	0.133
Molecule-Single	898.8	186.8	0.208	114.8	0.636
Molecule-Cluster	921.8	201.0	0.218	81.0	0.407

Table J2: Additional metrics for EC_{50} estimation and activity decision evaluation on the M2OR dataset presented in Figure 3. *Experimental error* is the error of repeated assays for a given protein-molecule pair and the performance lower bound. *Naïve* is the performance using the frequency statistics, i.e., $P(active|s, m) = \hat{p}_s\hat{p}_m$, where $\hat{p}_s$ and $\hat{p}_m$ are marginal probabilities (Lalis et al., 2024b). Standard deviation is given in parentheses. RMSLE is calculated on the intersection of the correctly predicted active pairs of ASMI-DR and DTA baselines. HyperAttentionDTI is abbreviated to HypAttnDTI and standard deviation is given in parentheses.

	Name	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
in vitro	Primary sc.	0.238	0.563	0.399		
	Secondary sc.	0.476	0.704	0.451		
	Experimental error				0.334	
i.i.d.	Naïve	0.365 (0.045)	0.386 (0.040)	0.470 (0.073)		
	Docking	0.215 (0.043)	0.447 (0.016)	0.647 (0.128)		
	MolTrans	0.590 (0.045)	0.590 (0.066)	0.857 (0.041)		
	HypAttnDTI	0.656 (0.029)	0.729 (0.030)	0.749 (0.034)		
	MAARDTI	<u>0.659</u> (0.024)	0.748 (0.025)	0.717 (0.027)		
	GNN-CLS	0.646 (0.026)	0.699 (0.035)	0.773 (0.030)		
	ASMI-Prob	0.640 (0.033)	0.748 (0.039)	0.699 (0.052)		
	Boltz-2	0.108 (0.033)	0.541 (0.117)	0.052 (0.006)	1.116 (0.056)	0.202 (0.050)
	DTIAM	0.658 (0.029)	0.646 (0.053)	0.854 (0.030)	1.149 (0.078)	0.535 (0.100)
	DeepDTAGen	0.592 (0.043)	0.631 (0.079)	0.757 (0.028)	1.287 (0.225)	0.352 (0.131)
	ProSmith	0.654 (0.030)	0.649 (0.041)	0.841 (0.007)	1.082 (0.060)	0.529 (0.063)
	ASMI-Reg	0.621 (0.048)	0.692 (0.047)	0.741 (0.032)	1.335 (0.412)	0.445 (0.142)
	ASMI-Cen	0.626 (0.029)	0.694 (0.044)	0.744 (0.025)	<u>0.888</u> (0.062)	<u>0.545</u> (0.051)
	ASMI-DR	**0.665** (0.011)	0.776 (0.024)	0.708 (0.027)	**0.683** (0.070)	**0.675** (0.071)

Table J3: Additional metrics for EC_{50} estimation and activity decision evaluation on the M2OR dataset presented in Figure 3. HyperAttentionDTI is abbreviated to HypAttnDTI and standard deviation is given in parentheses.

		Name	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
Sequence	Single	MolTrans	0.477 (0.082)	0.543 (0.109)	0.801 (0.065)		
		HypAttnDTI	0.480 (0.034)	0.600 (0.084)	0.684 (0.130)		
		MAARDTI	0.462 (0.090)	0.628 (0.085)	0.559 (0.100)		
		GNN-CLS	**0.525** (0.090)	0.652 (0.103)	0.669 (0.084)		
		ASMI-Prob	0.443 (0.057)	0.600 (0.083)	0.613 (0.057)		
		Boltz-2	0.136 (0.022)	0.883 (0.137)	0.036 (0.012)	1.139 (0.058)	0.255 (0.136)
		DTIAM	0.332 (0.042)	0.396 (0.061)	0.819 (0.110)	1.303 (0.227)	0.468 (0.121)
		DeepDTAGen	0.349 (0.061)	0.472 (0.088)	0.665 (0.170)	1.196 (0.259)	0.382 (0.027)
		ProSmith	0.426 (0.085)	0.461 (0.086)	0.844 (0.058)	1.160 (0.196)	0.497 (0.067)
		ASMI-Reg	0.398 (0.112)	0.506 (0.142)	0.690 (0.162)	1.493 (0.329)	0.228 (0.174)
		ASMI-Cen	0.355 (0.122)	0.485 (0.141)	0.589 (0.125)	1.163 (0.395)	0.423 (0.150)
		ASMI-DR	0.481 (0.031)	0.642 (0.038)	0.614 (0.095)	**0.718** (0.181)	**0.536** (0.195)
	Cluster	MolTrans	0.095 (0.062)	0.274 (0.107)	0.951 (0.057)		
		HypAttnDTI	0.232 (0.075)	0.436 (0.107)	0.399 (0.081)		
		MAARDTI	0.204 (0.109)	0.369 (0.154)	0.359 (0.153)		
		GNN-CLS	0.179 (0.047)	0.370 (0.095)	0.412 (0.112)		
		ASMI-Prob	0.207 (0.096)	0.405 (0.123)	0.377 (0.164)		
		Boltz-2	0.135 (0.066)	0.624 (0.233)	0.075 (0.054)	0.958 (0.166)	0.311 (0.324)
		DTIAM	0.210 (0.064)	0.301 (0.089)	0.708 (0.138)	2.328 (0.308)	0.275 (0.071)
		DeepDTAGen	0.210 (0.082)	0.348 (0.112)	0.481 (0.159)	1.347 (0.243)	0.005 (0.098)
		ProSmith	0.073 (0.098)	0.255 (0.133)	0.502 (0.059)	2.098 (0.110)	**0.428** (0.130)
		ASMI-Reg	**0.238** (0.123)	0.362 (0.113)	0.447 (0.162)	1.733 (0.558)	-0.048 (0.204)
		ASMI-Cen	0.136 (0.101)	0.319 (0.090)	0.487 (0.248)	1.223 (0.556)	0.248 (0.346)
		ASMI-DR	0.218 (0.043)	0.461 (0.162)	0.324 (0.131)	**0.951** (0.171)	0.015 (0.198)
Molecule	Single	MolTrans	0.588 (0.070)	0.553 (0.095)	0.838 (0.021)		
		HypAttnDTI	0.608 (0.073)	0.701 (0.072)	0.707 (0.095)		
		MAARDTI	0.524 (0.046)	0.642 (0.086)	0.602 (0.041)		
		GNN-CLS	0.577 (0.071)	0.660 (0.072)	0.707 (0.048)		
		ASMI-Prob	**0.614** (0.065)	0.702 (0.069)	0.716 (0.073)		
		Boltz-2	0.123 (0.054)	0.594 (0.108)	0.052 (0.018)	1.065 (0.065)	0.298 (0.132)
		DTIAM	0.570 (0.026)	0.547 (0.068)	0.839 (0.053)	1.422 (0.078)	0.381 (0.160)
		DeepDTAGen	0.579 (0.058)	0.599 (0.046)	0.766 (0.080)	1.206 (0.030)	0.326 (0.169)
		ProSmith	0.520 (0.052)	0.500 (0.083)	0.831 (0.071)	1.455 (0.166)	0.285 (0.081)
		ASMI-Reg	0.531 (0.054)	0.572 (0.086)	0.770 (0.089)	1.559 (0.259)	0.283 (0.088)
		ASMI-Cen	0.525 (0.040)	0.628 (0.042)	0.653 (0.051)	1.637 (0.541)	0.303 (0.148)
		ASMI-DR	0.593 (0.074)	0.663 (0.098)	0.738 (0.104)	**0.915** (0.085)	**0.487** (0.111)
	Cluster	MolTrans	0.434 (0.102)	0.513 (0.177)	0.768 (0.040)		
		HypAttnDTI	**0.482** (0.064)	0.620 (0.071)	0.591 (0.095)		
		MAARDTI	0.427 (0.012)	0.649 (0.056)	0.432 (0.057)		
		GNN-CLS	0.358 (0.109)	0.505 (0.150)	0.546 (0.150)		
		ASMI-Prob	0.458 (0.107)	0.653 (0.074)	0.502 (0.168)		
		Boltz-2	0.006 (0.058)	0.240 (0.269)	0.011 (0.012)	0.987 (0.100)	0.114 (0.079)
		DTIAM	0.412 (0.099)	0.439 (0.137)	0.769 (0.064)	1.087 (0.149)	0.281 (0.183)
		DeepDTAGen	0.391 (0.072)	0.503 (0.045)	0.582 (0.124)	1.116 (0.152)	0.240 (0.163)
		ProSmith	0.388 (0.066)	0.393 (0.093)	0.826 (0.059)	0.974 (0.103)	0.302 (0.173)
		ASMI-Reg	0.395 (0.082)	0.548 (0.151)	0.580 (0.130)	1.230 (0.322)	0.208 (0.207)
		ASMI-Cen	0.413 (0.100)	0.482 (0.100)	0.748 (0.132)	0.990 (0.153)	0.240 (0.180)
		ASMI-DR	0.398 (0.077)	0.572 (0.115)	0.526 (0.149)	**0.706** (0.031)	**0.345** (0.146)

J.2 INDEPENDENT EC50 EVALUATION FOR EACH MODEL ON M2OR

Figure 3, Table J2 and Table J3 report RMSLE calculated on the intersection of the correctly predicted active pairs across the models. Alternatively, in Table J4 we consider a standard scenario in practice when only a single model is applied and estimate the expected RMSLE on individual true positive activity predictions of each model.

Table J4: Evaluation of the EC_{50} estimation on true positive activity predictions of each model independently. Standard deviation is given in parentheses. [a]Due to the low MCC, the evaluation of Boltz-2 is done on all active pairs including the incorrectly classified pairs.

		Name	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
i.i.d.		Boltz-2	0.108 (0.033)	0.541 (0.117)	0.052 (0.006)	1.110[a] (0.037)	0.157[a] (0.049)
		DTIAM	0.658 (0.029)	0.646 (0.053)	0.854 (0.030)	1.516 (0.093)	0.408 (0.083)
		DeepDTAGen	0.592 (0.043)	0.631 (0.079)	0.757 (0.028)	1.417 (0.243)	0.338 (0.120)
		ProSmith	0.654 (0.030)	0.649 (0.041)	0.841 (0.007)	1.402 (0.111)	0.425 (0.087)
		ASMI-Reg	0.621 (0.048)	0.692 (0.047)	0.741 (0.032)	1.213 (0.135)	0.399 (0.122)
		ASMI-Cen	0.626 (0.029)	0.694 (0.044)	0.744 (0.025)	1.039 (0.032)	0.471 (0.031)
		ASMI-DR	**0.665** (0.011)	0.776 (0.024)	0.708 (0.027)	**0.708** (0.064)	**0.650** (0.061)
Sequence	**Single**	Boltz-2	0.136 (0.022)	0.883 (0.137)	0.036 (0.012)	1.045[a] (0.028)	0.128[a] (0.042)
		DTIAM	0.332 (0.042)	0.396 (0.061)	0.819 (0.110)	1.776 (0.192)	0.196 (0.104)
		DeepDTAGen	0.349 (0.061)	0.472 (0.088)	0.665 (0.170)	1.372 (0.161)	0.297 (0.075)
		ProSmith	0.426 (0.085)	0.461 (0.086)	0.844 (0.058)	1.693 (0.160)	0.240 (0.086)
		ASMI-Reg	0.398 (0.112)	0.506 (0.142)	0.690 (0.162)	1.543 (0.512)	0.150 (0.281)
		ASMI-Cen	0.355 (0.122)	0.485 (0.141)	0.589 (0.125)	1.512 (0.462)	0.254 (0.108)
		ASMI-DR	**0.481** (0.031)	0.642 (0.038)	0.614 (0.095)	**0.769** (0.156)	**0.470** (0.119)
	Cluster	Boltz-2	0.135 (0.066)	0.624 (0.233)	0.075 (0.054)	**1.069**[a] (0.058)	**0.222**[a] (0.113)
		DTIAM	0.210 (0.064)	0.301 (0.089)	0.708 (0.138)	2.711 (0.149)	0.053 (0.165)
		DeepDTAGen	0.210 (0.082)	0.348 (0.112)	0.481 (0.159)	1.946 (0.331)	-0.082 (0.144)
		ProSmith	0.073 (0.098)	0.255 (0.133)	0.502 (0.059)	2.524 (0.212)	0.174 (0.186)
		ASMI-Reg	**0.238** (0.123)	0.362 (0.113)	0.447 (0.162)	1.889 (0.481)	-0.145 (0.158)
		ASMI-Cen	0.136 (0.101)	0.319 0.090	0.487 (0.248)	1.791 (0.493)	-0.010 (0.234)
		ASMI-DR	0.217 (0.045)	0.461 (0.162)	0.324 (0.131)	1.168 (0.268)	0.040 (0.150)
Molecule	**Single**	Boltz-2	0.123 (0.054)	0.594 (0.108)	0.052 (0.018)	1.076[a] (0.080)	0.212[a] (0.079)
		DTIAM	0.570 (0.026)	0.547 (0.068)	0.839 (0.053)	1.672 (0.091)	0.341 (0.174)
		DeepDTAGen	0.579 (0.058)	0.599 (0.046)	0.766 (0.080)	1.267 (0.080)	0.326 (0.118)
		ProSmith	0.520 (0.052)	0.500 (0.083)	0.831 (0.071)	1.614 (0.110)	0.291 (0.108)
		ASMI-Reg	0.531 (0.054)	0.572 (0.086)	0.770 (0.089)	1.729 (0.347)	0.286 (0.152)
		ASMI-Cen	0.525 (0.040)	0.628 (0.042)	0.653 (0.051)	1.719 (0.502)	0.256 (0.117)
		ASMI-DR	**0.593** (0.074)	0.663 (0.098)	0.738 (0.104)	**0.920** (0.096)	**0.474** (0.116)
	Cluster	Boltz-2	0.006 (0.058)	0.240 (0.269)	0.011 (0.012)	0.981[a] (0.079)	0.096[a] (0.073)
		DTIAM	0.412 (0.099)	0.439 (0.137)	0.769 0.064	1.491 (0.208)	0.204 (0.144)
		DeepDTAGen	0.391 (0.072)	0.503 (0.045)	0.582 (0.124)	1.315 (0.191)	0.257 (0.124)
		ProSmith	0.388 (0.066)	0.393 (0.093)	0.826 (0.059)	1.513 (0.158)	0.139 (0.131)
		ASMI-Reg	0.395 (0.082)	0.548 (0.151)	0.580 (0.130)	1.561 (0.371)	0.154 (0.107)
		ASMI-Cen	**0.413** (0.100)	0.482 (0.100)	0.748 (0.132)	1.276 (0.264)	0.181 (0.135)
		ASMI-DR	0.398 (0.077)	0.572 (0.115)	0.526 (0.149)	**0.818** (0.154)	**0.298** (0.116)

J.3 INDEPENDENT K_D EVALUATION FOR EACH MODEL ON DTA BENCHMARKS

To have a fair evaluation, we report in Table 1 RMSLE calculated on the intersection of the correctly predicted active pairs across the models. In contrast, to estimate the expected performance when only

a single model is applied, we report in Table J5 RMSLE evaluated on individual true positives of each model.

Table J5: Individual K_d estimation evaluation on five i.i.d. splits. RMSLE is calculated on the individual true positive activity predictions of each model, where for the regression baselines, we consider pairs with the estimated $K_d > 7.94\mu$M as inactive. Standard deviation is given in parentheses.

	Model	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
DAVIS	DTIAM	0.566 (0.014)	0.543 (0.013)	0.939 (0.008)	0.736 (0.010)	0.669 (0.017)
	DeepDTAGen	0.642 (0.014)	0.691 (0.015)	0.816 (0.014)	0.788 (0.017)	0.621 (0.021)
	ProSmith	0.631 (0.010)	0.619 (0.013)	0.911 (0.007)	**0.688** (0.004)	**0.696** (0.009)
	ASMI-Reg	0.601 (0.048)	0.646 (0.063)	0.818 (0.030)	0.772 (0.057)	0.639 (0.023)
	ASMI-Cen	**0.685** (0.019)	0.818 (0.016)	0.722 (0.029)	0.772 (0.036)	0.637 (0.023)
	ASMI-DR	**0.685** (0.005)	0.772 (0.016)	0.775 (0.017)	0.713 (0.028)	**0.696** (0.014)
BindingDB	DTIAM	0.606 (0.026)	0.678 (0.021)	0.964 (0.005)	0.820 (0.020)	0.773 (0.010)
	DeepDTAGen	0.722 (0.014)	0.827 (0.018)	0.879 (0.009)	0.865 (0.029)	0.747 (0.014)
	ProSmith	0.607 (0.041)	0.678 (0.033)	0.966 (0.008)	**0.808** (0.026)	**0.780** (0.012)
	ASMI-Reg	0.446 (0.231)	0.620 (0.140)	0.946 (0.043)	0.926 (0.116)	0.699 (0.081)
	ASMI-Cen	0.727 (0.030)	0.873 (0.032)	0.823 (0.023)	0.923 (0.098)	0.694 (0.069)
	ASMI-DR	**0.745** (0.016)	0.839 (0.018)	0.890 (0.007)	0.834 (0.032)	0.774 (0.013)

J.4 Evaluation on drug-target interaction datasets

To validate our architectural choices in Section 5, we evaluate a concentration-free DTI version of the proposed architecture, ASMI-Prob, on two standard DTI benchmarks: KIBA (Tang et al., 2014) and DAVIS (Davis et al., 2011). We compare our architecture against state-of-the-art methods: MolTrans (Huang et al., 2020), HyperAttentionDTI (Zhao et al., 2021), GNN-CLS (Hladiš et al., 2023), and MAARDTI (Zhan et al., 2025).

As shown in Table J6, our architecture outperforms MolTrans, MAARDTI and GNN-CLS in AveP on both benchmarks. It lags behind HyperAttentionDTI by 2.1% and 1.9% on KIBA and DAVIS, respectively, with a superior precision, whereas HyperAttentionDTI exhibits a higher recall. Notably, the MCC of ASMI-DR trained by Algorithm 1 also surpasses all the DTI baselines on the DAVIS dataset. This result demonstrates that our architecture is not only effective for its primary OR-molecule activation task but also robust enough to compete with the state-of-the-art on standard DTI benchmarks.

Table J6: Performance on the KIBA and DAVIS drug-target interaction benchmarks. Standard deviation is given in parentheses.

	Model	AveP	Precision	Recall	MCC
DAVIS	MolTrans	0.798 (0.017)	0.593 (0.041)	0.857 (0.023)	0.584 (0.022)
	HyperAttentionDTI	**0.844** (0.008)	0.767 (0.014)	0.774 (0.016)	0.680 (0.006)
	MAARDTI	0.730 (0.023)	0.673 (0.020)	0.637 (0.031)	0.525 (0.023)
	GNN-CLS	0.744 (0.015)	0.699 (0.013)	0.666 (0.016)	0.564 (0.011)
	ASMI-Prob	0.828 (0.028)	0.790 (0.012)	0.680 (0.028)	0.639 (0.023)
	ASMI-Reg	0.817 (0.026)	0.646 (0.063)	0.818 (0.030)	0.601 (0.048)
	ASMI-DR	0.808 (0.015)	0.772 (0.016)	0.775 (0.017)	**0.685** (0.005)
KIBA	MolTrans	0.767 (0.026)	0.501 (0.032)	0.875 (0.009)	0.583 (0.030)
	HyperAttentionDTI	**0.819** (0.004)	0.718 (0.008)	0.770 (0.005)	**0.680** (0.005)
	GNN-CLS	0.704 (0.010)	0.610 (0.014)	0.683 (0.005)	0.555 (0.012)
	MAARDTI	0.715 (0.079)	0.708 (0.029)	0.667 (0.059)	0.615 (0.029)
	ASMI-Prob	0.802 (0.012)	0.738 (0.024)	0.655 (0.059)	0.628 (0.034)

J.5 EC_{50} ORDER EVALUATION

According to the primacy coding theory in olfaction (Wilson et al., 2017; Zwicker, 2019), the order of activation of olfactory receptors plays a pivotal role in the odour perception of a molecule. To assess the ability of the model to assign the rank, we evaluate the mean Spearman's rank correlation for individual molecules and proteins in Table J7. While the rank correlation per protein reaches a level comparable to the correlation among all the pairs (*ASMI-DR* in Table J7), the correlation per molecule is 40% lower in the i.i.d. case. Notably, the model achieves a higher ρ per molecule for novel molecules, but it fails to predict the order of activation per molecule when considering a new protein sequence.

Table J7: EC_{50} order evaluation. Spearman's ρ is the rank correlation between the EC_{50} estimated by ASMI-DR and the experimentally measured values. Rows *per molecule* and *per protein* correspond to the mean of correlation per molecule and per protein, respectively. The standard deviation of 5 runs is given in the parentheses.

		Name	MCC ↑	Precision ↑	Recall ↑	RMSLE ↓	Spearman's ρ ↑
i.i.d.		ASMI-DR	0.665 (0.011)	0.776 (0.024)	0.708 (0.027)	0.708 (0.064)	0.650 (0.061)
		per molecule					0.319 (0.225)
		per sequence					0.528 (0.052)
Sequence	**Single**	ASMI-DR	0.481 (0.031)	0.642 (0.038)	0.614 (0.095)	0.761 (0.150)	0.470 (0.119)
		per molecule					-0.134 (0.063)
		per sequence					0.464 (0.142)
	Cluster	ASMI-DR	0.218 (0.043)	0.461 (0.162)	0.324 (0.131)	1.170 (0.269)	0.040 (0.150)
		per molecule					0.167 (0.357)
		per sequence					0.151 (0.243)
Molecule	**Single**	ASMI-DR	0.593 (0.074)	0.663 (0.098)	0.738 (0.104)	0.920 (0.096)	0.474 (0.116)
		per molecule					0.244 (0.144)
		per sequence					0.290 (0.134)
	Cluster	ASMI-DR	0.398 (0.077)	0.572 (0.115)	0.526 (0.149)	0.818 (0.154)	0.298 (0.116)
		per molecule					0.246 (0.244)
		per sequence					0.150 (0.170)