Embracing assay heterogeneity with neural processes for markedly improved bioactivity predictions

Lucian Chan
Astex Pharmaceuticals
Cambridge, United Kingdom
lucian.chan@astx.com

Marcel Verdonk
Astex Pharmaceuticals
Cambridge, United Kingdom
marcel.verdonk@astx.com

Carl Poelking
Astex Pharmaceuticals
Cambridge, United Kingdom
carl.poelking@astx.com

Abstract

Predicting the bioactivity of a ligand is one of the hardest and most important challenges in computer-aided drug discovery. Despite years of data collection and curation efforts, bioactivity data remains sparse and heterogeneous, thus hampering efforts to build predictive models that are accurate, transferable and robust. The intrinsic variability of the experimental data is further compounded by data aggregation practices that neglect heterogeneity to overcome sparsity. Here we discuss the limitations of these practices and present a hierarchical meta-learning framework that exploits the information synergy across disparate assays by successfully accounting for assay heterogeneity. We show that the model achieves a drastic improvement in affinity prediction across diverse protein targets and assay types compared to conventional baselines. It can quickly adapt to new target contexts using very few observations, thus enabling large-scale virtual screening in early-phase drug discovery.

1 Introduction

The primary aim of drug discovery is to design compounds that are safe for the patient and efficacious against the disease. During preclinical development, efficacy is usually equated with potency: A drug that is a potent inhibitor of a target protein is assumed to be effective against the associated disease. Hence building potency while maintaining a safe chemical profile is the key objective in early-phase drug discovery. Models that can predict the binding affinity as a measure of potency of a small molecule against a protein target are therefore the subject of intense research efforts. Existing physics-based approaches such as free-energy perturbation, however, are computationally extremely demanding, and despite their computational cost suffer from potentially significant systematic errors. Data-driven and deep-learning-based alternatives could thus help reduce the computational burden and help identify candidate molecules across vast chemical spacess (1; 2; 3; 4; 5).

Among such data-driven techniques, we conventionally distinguish between *ligand-based* (6; 7; 8; 9; 10) and *structure-based* approaches (11; 12; 13; 14) that typically produce models that are either *local* (i.e., specific to a certain protein system) or *global* in scope. One key issue in the construction of these models is that the available bioactivity data are globally sparse and locally heterogeneous. As a result, models will struggle to identify relevant patterns across apparently *unrelated* assays (unrelated because they might target very different proteins); and, also, that they struggle to reconcile apparent inconsistencies among supposedly *related* assays (related because they might target similar proteins or assay similar compounds). For the latter, we use the term *heterogeneous* to describe the case where two seemingly equivalent assays measure systematically different structure-activity relationships (SARs). This heterogeneity – or *between-assay* variability – is therefore not caused by experimental uncertainty – the *within-assay* variance (15; 16). Instead, it is

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

the result of differences in the experimental approach, assay type, assay conditions, and unobserved or undeclared hidden variables that affect the assay outcome: Consider, for example, cellular vs non-cellular and direct-binding vs functional assays; variations in pH, temperature, buffer composition or protein concentration; batch effects, solubility issues and non-specific binding; differences in the sensitivities and dynamic ranges of the experimental setups, or differences in the data analysis. Such differences in conditions and assay types are, however, routinely ignored when collating data across assays to create a single dataset per protein: a form of point-wise aggregation that can cause severe issues downstream in the modelling process.

2 This Work

Due to the heterogeneity of bioactivity data, we propose that the data aggregation process needs to be learnt implicitly by the predictive model itself, and that the model must be agnostic of the exact source of the heterogeneity. Here we show that this objective can be met with a hierarchical meta-learning model that incorporates assay heterogeneity while improving data efficiency. The model design reflects the fact that in a drug-discovery setting, we routinely find ourselves faced with a few-shot learning problem, as at best a few hundred to thousand molecules are made and tested during the typical life cycle of a discovery project. Importantly, the model is able to infer an assay-specific SAR from only a small number of observed data points in a few-shot manner, and is therefore able to 'translate' results from one assay into the context of another. This assay-specific approach differs conceptually from previous meta-learning attempts and produces, as we show, a drastic improvement in predictive performance. In summary, our main contributions are as follows:

- We illustrate that assay heterogeneity is a key cause of poor learning behaviour and model performance.
- We propose a hierarchical meta-learning technique (MetaBind) that simultaneously addresses data sparsity and heterogeneity in bioactivity modelling.
- We evaluate MetaBind on a curated bioactivity dataset, including a paired-assay split that measures directly the model's capability to infer assay-dependent SARs.

3 Related Work

Molecular multi-task learning. Multi-task models have a long-standing tradition in molecular-property prediction, and have also been developed for bioactivity modelling (17; 18; 19). These models generally treat each assay as a task and model all tasks jointly. This formulation avoids the assay heterogeneity effect introduced in point-wise aggregation, and explicitly exploits the correlation between assays to predict the bioactivity of new protein-ligand pairs. These approaches, however, typically require a moderate number of observations per task (> 50) and some degree of compound overlap among the different assays, which renders them impractical to apply to low-data protein targets.

Meta-learning. Meta-learning is an umbrella term for machine-learning approaches that aim to help a model adapt to new tasks using information on previous tasks. Various studies have applied meta-learning principles to molecular property prediction (20; 21; 22) and molecular optimization (23) in low-data regimes. These meta-learning techniques can be broadly classified into model-based, optimization-based (24) and metric-based (25) approaches. Model-based techniques aggregate information from previous tasks to extrapolate to new tasks and contexts. Optimization-based approaches specialise a meta-learner on a new task using, e.g., gradient-based optimization on a support set. Metric-based approaches construct a similarity metric across tasks and use embedding-based queries to contextualize a support set. These proposed models were primarily developed for predictions on new targets or endpoints in a low-data regime, whereas assay heterogeneity was ignored.

Neural Processes. Neural Processes (NPs) (26) learn a stochastic process over functions as solution to a supervised learning problem. They thus combine the adaptability of neural networks with the characteristics of Gaussian Processes. The neural process is constructed in two stages. First, it learns the statistics of a generic domain from a large sample set without committing to a specific learning task. Second, using these domain-wide statistics, it constructs a distribution over functions

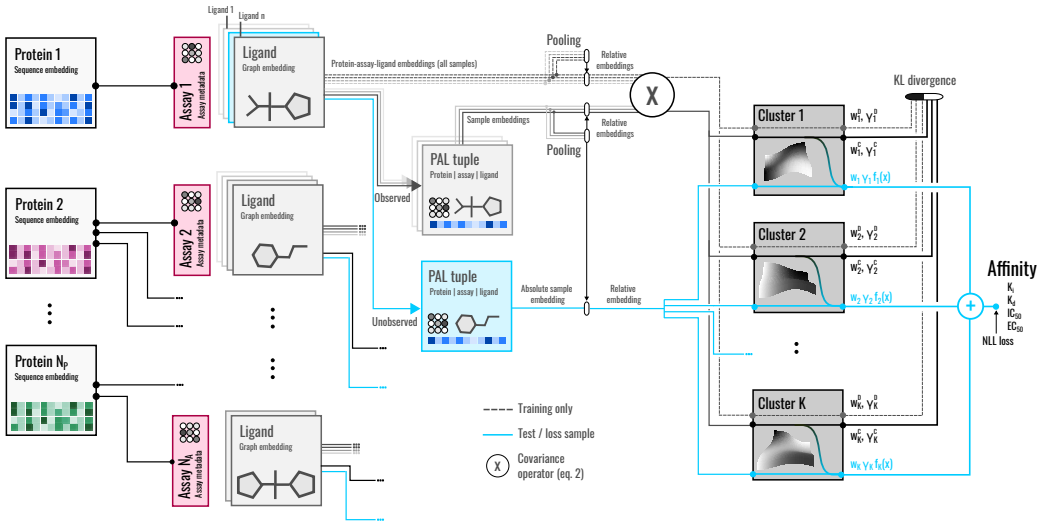

Figure 1: **Illustration of the MetaBind approach.** The model constructs an assay-specific SAR using a local support set of per-assay observations to predict affinities for unobserved protein-ligand pairs. Each assay is treated as a task, and assays are clustered into K covariance groups. The subset of observed activities and the associated protein-ligand pairs are used to construct the SAR representation. At training time, this subset is randomly sampled from the available data on a per-assay basis, with the model trying to reconstruct the covariance structure associated with the full set of observations (dashed lines up top) based on a Kullback-Leibler loss term. The aggregation function, which is here based on a cross-covariance operator, as well as the protein and ligand embeddings are implemented using convolutional or (for ligands) graph-convolutional neural networks. Note that the assay-specific SAR is based, first and foremost, on the covariance structure inferred from the observed bioactivities of the context set, in line with the notion that all assay condition variables are implicitly encoded into the bioactivities themselves.

for a specific task from a small support set. An NP variant also includes a global latent variable, z, to account for the uncertainty in the predictions of the output for given input pairs (x, y) (27). Its application to molecular-property prediction has been explored recently (28; 29). We note that standard NPs only model data from a single stochastic process and are thus designed to infer each task independently. There have been various efforts to extend NPs to a multi-task setting (30)

4 Methods and Data

The meta-learning approach presented in this section models assay heterogeneity by exploiting the underlying correlation structure across diverse sets of assays. The resulting model, dubbed MetaBind, applies meta-learning at the assay level (each experiment thus constituting a separate task) to learn a local assay-specific SAR; it is a clustered multi-task neural process that can be viewed as an extension of standard neural processes (26).

4.1 MetaBind formalism

Conventionally, to construct a neural process, task-specific samples, denoted as context data, C, are collected in order to infer the statistics of the associated domain. Based on these domain-wide statistics learned from the context, the model predicts the distribution $p(y|x, C)$ of a target output y for a given input x. In our case, the context data consists of the ligand structures and read-outs from a given assay. The learned statistics not only capture information on the underlying SAR, but also implicitly the experimental conditions. Using this framework, ligand affinities and uncertainties in the predictions can therefore be inferred independently for each assay, in contrast to existing models where the assay heterogeneity is neglected during training. The statistics generated from the context data is critical as it has to encapsulate the functional form of the SAR and be rich

enough to identify variability across assays. Nevertheless, despite their flexibility, scalability and well calibrated uncertainty estimation, standard neural processes are designed to model data from only a *single* stochastic process, which render them unsuitable for bioactivity data collected using various techniques (e.g., biophysical, biochemical or cell-based assays), with various endpoints (e.g., K_i, K_d, IC_{50} and EC_{50}) and variable experimental conditions.

A simple way of incorporating heterogeneity among related or equivalent assays would be to estimate an affinity offset from protein-ligand pairs that are shared across experiments (i.e., calibration via reference compounds). Such pairs are, however, few and far between within mixed-origin databases such as ChEMBL. Therefore, instead of estimating the offset directly – and thus limiting the model to constant-offset heterogeneities – the idea is to model the *relative* change within assays while incorporating information from related experiments in a mechanistically agnostic and functionally flexible manner. This is why here we introduce a clustered multi-task neural process (Fig. 1) that represents the bioactivity as a linear combination of a learnt reference affinity value $\tilde{y}^{\text{ref}}$, and K *independent* stochastic processes f_k with scaling factor γ that model the relative SAR within the assay. I.e., we use the ansatz

$$y_{ia} = \; y_a^{\text{ref}} + \sum_{k=1}^{K} w_{ka}\big(\gamma_a f_k(x_{ia} - x_a^{\text{ref}})\big) + \epsilon_{ia}, \qquad \text{s.t.} \sum_{k=1}^{K} w_{ka} = 1, \tag{1}$$

where x_{ia} is the representation of protein-ligand pair i, and y_{ia} is the associated bioactivity relationship for assay a; ϵ_{ia} is an error term that follows a Gaussian distribution with zero mean and constant variance. The scales γ_a and weights w_{ka} are learnable functions (see the Appendix for details).

This formulation assumes that the assays can be clustered into a small number of covariance groups, and that the SAR within each group is defined by a mean function f_k and assay-dependent scaling factor γ_a, as well as weight functions $(w_{1a}, ..., w_{ka})$ that quantify the relevance of each assay in the respective clusters. This functional structure makes the key difference to existing models, as assays are not aggregated in a point-wise fashion before the model is constructed, but instead aggregated implicitly on the fly in a manner learnt by the model itself.

Note that the cluster functions f_k operate on a relative embedding $\tilde{x} = x_{ia} - x_a^{\text{ref}}$ to emphasise the variability within an assay relative to a reference calculated over the context C,

$$(x_a^{\text{ref}}, y_a^{\text{ref}}) = \left(\frac{\sum_i^{N_a^C} x_i}{N_a^C}, \frac{\sum_i^{N_a^C} y_{ia}}{N_a^C} \right). \tag{2}$$

To effectively aggregate functionally related assays, one could model the joint distribution of $(\tilde{x}, \tilde{y})$, *i.e.,* $P_{\tilde{x}\tilde{y}}$, through mean embedding of concatenated samples, $\phi(\tilde{x}, \tilde{y})$. Here we instead represent this joint distribution using the cross-covariance operator C_{XY}(31) over the context C:

$$C_{XY}^a = \frac{1}{||C_a||} \sum_{i \in C_a} \phi_x(\tilde{x}_{ia}) \otimes \phi_y(\tilde{y}_{ia}) = \frac{1}{||C_a||} \Phi_x^T \Phi_y^a \in \mathbb{R}^{h \times d}, \tag{3}$$

where $\otimes$ denotes the outer product, $\phi_{\tilde{x}}$ and $\phi_{\tilde{y}}$ are feature maps for covariates and dependent variables, respectively; h and d are the output dimensions of $\phi_{\tilde{x}}$ and $\phi_{\tilde{y}}$. This cross-covariance is then flattened to parameterize the learned statistics via the weights $w_{ka} = w_k(C_{XY}^a)$ and scale parameter $\gamma_a = \gamma(C_{XY}^a)$. Conceptually, similarity of the SARs of two assays a and b implies similar statistics, $P_{\tilde{x}\tilde{y}}^a \simeq P_{\tilde{x}\tilde{y}^b}$, and thus similar functional parameters w_k and γ.

This multi-modal approach with relative embeddings ensures that the neural process can capture non-trivial data heterogeneities. In the simple case that two assays differ by a mean shift caused by, for example, an annotation error that substituted nanomolar with micromolar units, the change is fully encapsulated in the reference affinity y^{ref}, whereas the induced shape parameters are unaffected, as the covariance structure C_{XY} remains unchanged. Concept shifts in the SAR, by contrast, would trigger adjustments in $(\mathbf{w}, \gamma)$, in addition to potential changes in the reference itself.

Proteins and ligands are featurized via their sequence and molecular graphs, respectively. Please refer to the Supplementary Information (SI) for the featurization and encoding of protein-ligand pairs, the neural networks used to represent the shape parameters and cluster functions, and training protocol.

4.2 Data

We use bioactivity data from ChEMBL30 (32) for model training and evaluation. We considered binding assays with at least 15 measured bioactivities ($K_i/K_d/IC_{50}/EC_{50}$). To properly evaluate the model performance under assay heterogeneity with known ground truth, we introduce a paired-assay split: The test set of this split consists of 100 assays pairs where the partners share identical target proteins together with a series of at least fifteen ligands. This split is thus suitable to test directly the model's ability to infer an assay-specific SAR. Additionally, a chronological split assesses the model's performance for unseen ligands and protein targets.

Please refer to the Appendix and SI for further details on the construction of these splits, as well as on data preprocessing, filtering and curation.

4.3 Evaluation

We evaluate the model performance at the task (*i.e.,*, assay) level using a task mean squared error (T-RMSE) and mean absolute error (T-MAE):

$$\text{T-RMSE}_a = \sqrt{\frac{1}{N_a}\sum_{i=1}^{N_a}(\hat{y_{ia}} - y_{ia})^2}, \qquad \text{T-MAE}_a = \frac{1}{N_a}\sum_{i=1}^{N_a}|\hat{y}_{ia} - y_{ia}|, \tag{4}$$

with task $a \in \{1, ..., n\}$, observations $i \in \{1, ..., N_a\}$, and N_a the number of observations for task a; $\hat{y}$ is the predicted binding affinity (or bioactivity), y is the observed binding affinity (or bioactivity); $|\cdot|$ indicates the absolute value. We additionally use the Pearson's correlation coefficient r to assess the (global) correlation between the predicted and observed affinities.

5 Results

MetaBind is designed to handle assay heterogeneity in a functionally agnostic manner: I.e., it does not make any specific assumptions how the SARs of two related assays differ. We illustrate this aspect first by showcasing how the model behaves for different heterogeneity classes. We will then proceed with a broader validation that indicates performance levels in a semi-prospective or prospective setting relative to conventional baselines.

5.1 Illustration of heterogeneity effect on model predictions

We distinguish between three different grades of SAR heterogeneity: Simple offsets, offsets combined with rescaling, and fundamental conceptual shifts in the SAR. Fig. 2 shows data for three different assay pairs that fall within one of these heterogeneity classes. Two assay pairs (Fig. 2a-b) display distinct SARs despite the partners' sharing the same protein-ligand series; the third pair (Fig. 2c) compares experiments for the wild-type versus a mutated variant of the same target protein. The SARs for the first assay pair (panel a) are strongly correlated with a Pearson correlation coefficient $r = 0.93$, but a sizeable systematic offset of approximately 0.78 log units that might be the effect of human plasma that is present in one assay (assay A) but not the other (assay B).

The second pair (panel b) compares a cell-based (C) and biochemical assay (D) for the same protein. The SARs are essentially uncorrelated with $r = 0.26$, raising the question whether there is any meaningful information exchange possible between these assays. It is pairs like these that can cause severe issues during point-wise data aggregation, as the data samples are clearly drawn from two different distributions, and any agreement along the diagonal must be interpreted as coincidental.

The third pair (panel c) compares the SARs for two assays of different variants of the same protein: Assay E is a cell-based assay of wild-type B-Raf, whereas assay F is a biochemical assay of a singly mutated variant (V600E, which is a key mutation in various cancer types). Although the measured bioactivities are clearly correlated with $r = 0.63$, the SARs differ in terms of both scaling and offset. It is then impossible to determine without in-depth analysis what role the mutation plays in causing this difference, compared to the role of the variability in assay type, or the effect of a random heterogeneity due to unobserved or undeclared condition variables.

So how do machine-learning models fare at predicting this variability in SAR? First, consider a standard graph-convolutional neural network (GCNN) that is trained on ChEMBL data preprocessed

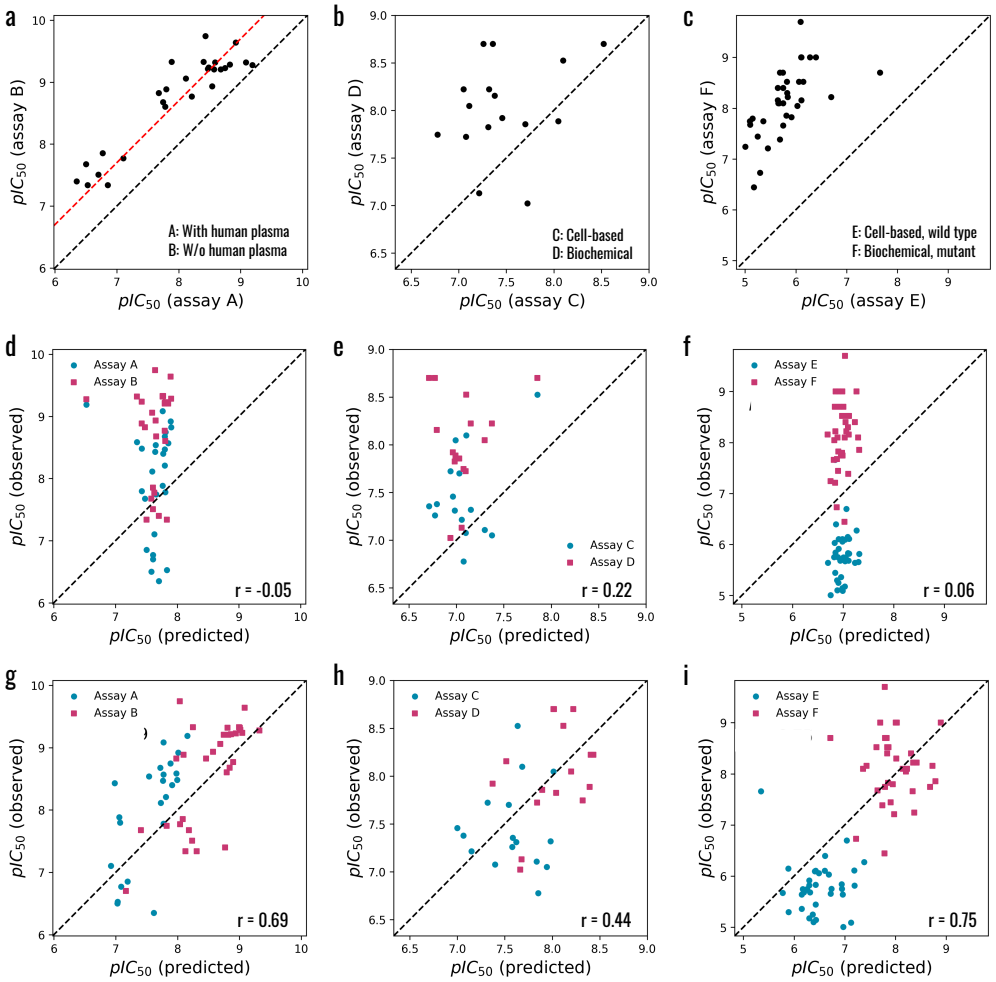

Figure 2: **Observed and predicted heterogeneities in structure-activity relationships.** (a-c) Correlation of the bioactivity read-out for congeneric series of three different assay pairs. The pairs (A,B) and (C,D) share the same protein variants, but differ in assay type and conditions. The assays of pair (E,F) differ in protein variant, with assay E investigating the wild type, and F a singly mutated variant thereof. (d-f) Baseline predictions for the series of assays A-F obtained with a standard graph-convolutional NN. The baseline reproduces neither the activity trends within the assays, nor between the partners. (g-i) Affinity predictions obtained with MetaBind, demonstrating improved performance and at least partial reconstruction of the heterogeneity for all three pairs.

using point-wise aggregation rules. Expectedly, this simple baseline performs extremely poorly on this type of problem (see Fig. 2d-f): Not only does the model ignore differences in SAR, but it also fails to produce any discernable correlation between its predictions and the experimental affinities *per assay*. Furthermore, its predictions tend to fall into a very narrow affinity range. Such behaviour can be caused by the learning dynamics of neural networks, which tend to produce smooth functions over the input space. As a result, small perturbations due to, e.g., minor changes in a lead series or single-point mutations of a protein, fail to produce a significant change in the predicted SAR. This behaviour is aggravated further if the functional form of the predictor remains static, which is the case for conventional GCNNs.

The MetaBind model proves to be significantly more flexible in this regard (Fig. 2g-i): Even though the intra-assay correlation of its predictions with the experimental observations is rather modest, the mean absolute errors for the different assays range between 0.41 and 0.85 pK units. Importantly, the

	Paired (biochemical)		Chronological (biochemical)	
	Task RMSE	Task MAE	Task RMSE	Task MAE
Baseline	1.125 (0.477)	0.981 (0.465)	1.150 (0.477)	1.008 (0.480)
Baseline (local)	1.102 (0.462)	0.960 (0.450)	1.129 (0.464)	0.988 (0.466)
MetaBind	**0.794 (0.251)**	**0.663 (0.231)**	**0.828 (0.270)**	**0.689 (0.243)**

	Paired (cell-based)		Chronological (cell-based)	
	Task RMSE	Task MAE	Task RMSE	Task MAE
Baseline	1.160 (0.493)	1.023 (0.500)	1.171 (0.497)	1.030 (0.498)
Baseline (local)	1.137 (0.484)	1.001 (0.487)	1.152 (0.485)	1.012 (0.487)
MetaBind	**0.810 (0.236)**	**0.672 (0.210)**	**0.854 (0.282)**	**0.712 (0.257)**

	Paired (all)		Chronological (all)	
	Task RMSE	Task MAE	Task RMSE	Task MAE
Baseline	1.110 (0.443)	0.974 (0.436)	1.131 (0.468)	0.992 (0.469)
Baseline (local)	1.090 (0.436)	0.955 (0.430)	1.113 (0.457)	0.974 (0.459)
MetaBind	**0.763 (0.234)**	**0.632 (0.208)**	**0.803 (0.252)**	**0.668 (0.227)**

Table 1: **MetaBind vs baseline performance metrics.** The baseline models are GCNNs constructed using pre-aggregated training data: *Baseline* does not use any additional test data during predictions, as opposed to *Baseline (local)*, which performs a local update of the model given the assay context data. For each test setting (paired and chronological), we distinguish between biochemical and cell-based data partitions. MetaBind outperforms both baselines by a sizeable margin across all splits and test scenarios. Note the value in brackets, which indicates the standard deviation of the per-assay RMSEs and MAEs (not the error of the mean, which is far smaller).

ligand heterogeneity, i.e., the difference in affinity for the same ligand i but different assays (a, b),

$$\Delta_{ia,ib} = y_{ib} - y_{ia}, \tag{5}$$

is reasonably well predicted with mean absolute errors (MAEs) of 0.35, 0.48 and 0.89 pK units for assay pairs (A,B), (C,D) and (E,F), respectively. These values should be compared to the mean absolute deviation of the measured ligand heterogeneities (which are observable here because of the joint ligand set) of 0.78, 0.68 and 2.33. This suggests that the model is able to capture the heterogeneity at least partially. Analysis of the functional parameters (cluster weights w and scales γ) gives further insight into the type of heterogeneity predicted by the model. E.g., for assay pair (A,B), the difference in predicted scaling is insignificant, $|\gamma_A - \gamma_B| \approx 0.005$, as is expected for a constant-offset heterogeneity. Meanwhile, the ranks of the cluster weights w are identical within each pair, illustrating that the partners are predicted to fall within a similar SAR covariance class, despite perhaps what the experimental read-out from assay pair (C,D) suggests.

5.2 Evaluation and benchmark on ChEMBL30

These examples underline why 'neural' data aggregation as implemented by MetaBind is expedient. For a broader validation of the approach, we again use ChEMBL30 data, this time to set up two different benchmark settings: A paired assay split and a chronological split. The former defines a test set of 190 assay pairs where the assays of each pair target the same protein, share at least 15 ligands, and have an intra-assay mean absolute deviation of at least 0.5 pK units. The latter uses a time split to divide the ChEMBL data into a training set (pre-2016) and a test set (post-2016), where the latter is made up of 178 unseen protein targets. Among the paired assays of the test set, only 34 exhibit a correlation $r > 0.7$, despite the Pearson correlation coefficient being invariant to constant offsets or rescaling of the SAR. See the SI for further details on the splitting procedure and dataset properties.

The paired-assay split is designed to test the model's ability to infer the SAR heterogeneity, whereas the chronological split measures how well the model generalizes to new assay contexts. We include two baseline models as reference: First, a conventional GCNN – denoted *Baseline* – trained on pre-aggregated data, and a variant thereof that uses gradient updates on context data from the test

set – denoted *Baseline (local)*. The latter enables a fair comparison with MetaBind, as additional context data is used in an optimization-based meta-learning-style approach. Because MetaBind infers an assay-specific SAR from a context support set, a direct comparison with existing bioactivity and meta-learning models is not possible for conceptual reasons.

The results of the comparison and MetaBind's test performance are summarised in Table 1 and Fig. S4a-b of the SI. We find that MetaBind outperforms the two baselines by a sizeable margin across all splits and metrics. The metrics we use are the assay-level task RMSEs and MAEs as defined above. For the test settings explored here, the mean task MAEs measured for the baselines range between 0.955 to 1.171, compared to 0.632 and 0.854 for MetaBind. For each of the two test scenarios (paired and chronological split) we additionally investigate different partitions of the ChEMBL data according to assay type (biochemical, cell-based, any). Cell-based assays capture the biological context more faithfully, but are a potentially only indirect measure of ligand activity against a protein target, less reproducible, and more susceptible to off-target effects. It is nevertheless reassuring that models built exclusively on cell-based data perform almost on par with those trained exclusively on biochemical data. Furthermore, despite significant differences in chemical- and protein-space coverage of the biochemical and cell-based data partitions, MetaBind achieves a noticeable gain in performance when trained on a combined dataset. This is likely due to its enhanced ability to model the additional heterogeneity introduced into the data upon aggregation of diverse assay types. Overall, the good metrics and healthy correlation plots (Fig. S4a-b) obtained with MetaBind indicate that the framework successfully constructs expressive SARs from local context data. The locally optimized baseline, by contrast, achieves only a very minor improvement over the standard GCNN.

Finally, we stress that estimating offsets, scalings or other heterogeneities across two assays without duplicate compounds is typically an intractable problem. For the paired-test setting, however, we have the experimental ground truths for the heterogeneities available as by construction. This means that we can directly compare the predicted per-ligand shifts $\Delta_{ia,ib}$ to their experimental values. The correlation (Fig. S4c) is excellent, with a Pearson r of 0.84 and a mean absolute deviation of 0.56 pK units. This is particularly encouraging as only five context observations were randomly selected from each assay to be supplied to the model, and not necessarily the same five within each pair. This indicates that the covariance structure $C_{\tilde{X}\tilde{Y}}$ induced by the context set is robust. Still, if the signal-to-noise ratio is extremely low, the cluster assignment should be expected to become unstable, resulting in conservative predictions biased towards the sample mean. Potential solutions for this noisy regime could be to either increase the size of the context set, or provide additional metadata, for instance on binding sites or assay conditions.

6 Conclusions

Here we proposed a meta-learning formalism, MetaBind, that infers local, assay-specific SARs for protein-ligand affinity modelling by harnessing both local prior knowledge and global, diverse bioactivity data. The model enables a form of neural data aggregation to address the issue of assay heterogeneity: systematic differences in SAR produced by equivalent assays which, if unaccounted for, can have a severe and deleterious effect on model performance. We show that MetaBind produces exceptional results in different benchmark settings and in a few-shot manner, far outperforming conventional baseline models trained on pre-aggregated data, even when these are optimized for a specific assay context. MetaBind is thus able to adapt quickly to new target proteins, ligands, and assays, rendering it easily applicable to a drug-discovery setting.

Testing on paired assays highlights that the formalism is able to reconstruct in a robust way the heterogeneity even when presented with only minimal support data from each assay partner. Importantly, MetaBind does not assume a fixed functional form for the heterogeneity (such as constant offsets or rescalings), does not rely on knowledge of its causal origin, or on duplicate compounds shared across assays.

Data heterogeneity is generally considered an issue that affects all major (bio-)chemical databases due to their mixed origin, diverse content and almost necessarily incomplete annotation. We therefore believe that formalisms such as MetaBind, which are based on an implicitly learnt, 'smart' data aggregation function, can prove useful in constructing transferable, performant models using community data not just for bioactivity modelling, but also chemical reactivity, ADME-Tox predictions and beyond.

7 Acknowledgements and Disclosure of Funding

LC acknowledges funding from Astex through the Sustaining Innovation Postdoctoral Programme. We thank Chris Murray, Davide Branduardi, Lisa Ronan, Tugce Oruc, Rudolfs Petrovs, Caroline Richardson and the anonymous reviewers for their valuable comments.

References

[1] Sadybekov, A. A. *et al.* Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022). URL https://www.nature.com/articles/s41586-021-04220-9.

[2] Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019). URL https://www.nature.com/articles/s41586-019-0917-9.

[3] Warr, W. A., Nicklaus, M. C., Nicolaou, C. A. & Rarey, M. Exploration of ultralarge compound collections for drug discovery. *Journal of Chemical Information and Modeling* **62**, 2021–2034 (2022). URL https://pubs.acs.org/doi/10.1021/acs.jcim.2c00224.

[4] Enamine REAL Space. https://enamine.net/library-synthesis/real-compounds/real-space-navigator (2022).

[5] GalaXi Space. https://www.labnetwork.com/frontend-app/p/#!/library/virtual (2022).

[6] Öztürk, H., Özgür, A. & Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018). URL https://academic.oup.com/bioinformatics/article/34/17/i821/5093245.

[7] Nguyen, T. *et al.* GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2020).

[8] Daga, A. *et al.* LEP-AD: Language Embedding of Proteins and Attention to Drugs predicts drug target interactions. In *ICLR MLDD workshop* (2023).

[9] Lee, E., Yoo, J., Lee, H. & Hong, S. Metadta: Meta-learning-based drug-target binding affinity prediction (2022). URL https://openreview.net/forum?id=yzlif16IASM.

[10] Wallach, I. & Heifets, A. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of Chemical Information and Modeling* **58**, 916–932 (2018).

[11] Volkov, M. *et al.* On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of Medicinal Chemistry* **65**, 7946–7958 (2022). URL https://pubs.acs.org/doi/10.1021/acs.jmedchem.2c00487.

[12] Yan, J. *et al.* Multi-task bioassay pre-training for protein-ligand binding affinity prediction (2023). URL http://arxiv.org/abs/2306.04886. ArXiv:2306.04886 [cs, q-bio].

[13] Jones, D. *et al.* Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of Chemical Information and Modeling* **61**, 1583–1592 (2021). URL https://pubs.acs.org/doi/10.1021/acs.jcim.0c01306.

[14] Kyro, G. W., Brent, R. I. & Batista, V. S. Hac-net: A hybrid attention-based convolutional neural network for highly accurate protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling* **63**, 1947–1960 (2023). URL https://pubs.acs.org/doi/10.1021/acs.jcim.3c00251.

[15] Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The experimental uncertainty of heterogeneous public Ki data. *J. Med. Chem.* **55**, 5165–5173 (2012).

[16] Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of mixed IC50 data – a statistical analysis. *PLOS ONE* **8**, 1–12 (2013).

[17] Martin, E. J. *et al.* All-assay-max2 pqsar: Activity predictions as accurate as four-concentration ic50s for 8558 novartis assays. *Journal of Chemical Information and Modeling* **59**, 4450–4459 (2019).

[18] Whitehead, T. M., Irwin, B. W. J., Hunt, P., Segall, M. D. & Conduit, G. J. Imputation of assay bioactivity data using deep learning. *Journal of Chemical Information and Modeling* **59**, 1197–1204 (2019).

[19] Pentina, A. & Clevert, D.-A. Multi-task proteochemometric modelling. *ChemRxiv* (2022).

[20] Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Central Science* **3**, 283–293 (2017).

[21] Nguyen, C. Q., Kreatsoulas, C. & Branson, K. M. Meta-learning initializations for low-resource drug discovery. In *ICML Workshop on Graph Representation Learning and Beyond (GRL+)* (2020).

[22] Pappu, A. & Paige, B. Making graph neural networks worth it for low-data molecular machine learning. In *NeurIPS Machine Learning for Molecules Workshop* (2020).

[23] Wang, J., Zheng, S., Chen, J. & Yang, Y. Meta learning for low-resource molecular optimization. *Journal of Chemical Information and Modeling* **61**, 1627–1636 (2021).

[24] Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 1126–1135 (2017).

[25] Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. In Guyon, I. *et al.* (eds.) *NIPS*, vol. 30 (2017).

[26] Garnelo, M. *et al.* Neural processes. *ArXiv* **abs/1807.01622** (2018).

[27] Kim, H. *et al.* Attentive neural processes. In *International Conference on Learning Representations* (2019).

[28] Lee, E., Yoo, J., Lee, H. & Hong, S. MetaDTA: Meta-learning-based drug-target binding affinity prediction. In *ICLR Machine Learning for Drug Discovery* (2022).

[29] Garcia-Ortegon, M., Bender, A. & Bacallado, S. Conditional neural processes for molecules. In *Machine Learning in Structural Biology* (2022).

[30] Kim, D., Cho, S., Lee, W. & Hong, S. Multi-task processes. In *International Conference on Learning Representations* (2022).

[31] Gretton, A. Notes on mean embeddings and covariance operators (2020).

[32] Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930—D940 (2019).

[33] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

[34] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2017).

[35] Wang, M. *et al.* Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315* (2019).

[36] RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

[37] Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

Appendix

A Data

ChEMBL 30 was used throughout the study. We considered binding assays that are assigned to a single protein target with a ChEMBL confidence score of at least 9 (the maximum score). We included four endpoints: K_i, K_d, IC_{50} and EC_{50}. The target species is *homo sapiens*. Here we do not consider censored activity data, i.e., we remove data samples with inequality labels $<$ or $>$. Moreover, we exclude assays that report identical activity values for more than five compounds, or those with an activity range of less than 0.5 pK units. Among the remaining set, we include only assays with more than ten exact measurements, and consider only ligands with heavy-atom counts ≤ 50 and the following heavy-atom types: C, N, O, F, P, S, Cl, Br, I. The resulting datasets consists of 6968 assays, with a total of 150,159 measured affinities across 100,279 unique ligands and 830 distinct target proteins. Note that we only aggregate duplicates *within* assays, but keep duplicates across assays separate. See the SI for a more detailed data summary.

we constructed a test set of assays pairs where the partners share identical target proteins together with a series of at least fifteen ligands. Additionally, we ensured that the mean absolute deviation of the bioactivities within the individual assays is 0.5 log units or larger. We thus identified 100 assay pairs formed from among 190 unique assays as test set, and denoted the corresponding split as the *paired-assay* split. The affinities of the assay partners are not generally strongly correlated, see the SI for details.

To assess how the models adapt to new assays and targets, we split the assays chronologically, with assays prior to 2016 entering the training set, assays of 2016 or later the test set. This results in a test set of 178 unseen target proteins. The median ligand similarity between training and testing set is 0.32. See further details in the SI.

B MetaBind architecture

We employed a clustered multitask neural process (CMTNP, eq. 1) to model the protein-ligand bioactivity because of its flexibility, scalability and capacity to produce well calibrated predictions. Derived from context data, the assay representation plays a crucial role in our framework as it provides the basis for how the model learns to aggregate data while constructing a local SAR. We use the cross-covariance operator in eq. 3 to model the functional dependency of the SAR as probed by the different protein-ligand pairs of an assay, *i.e.,* $P_{\tilde{X}\tilde{Y}}$. To infer appropriate scaling factors, the assay representation is normalised by the variance of the covariates. The full representation thus assumes the form

$$x_a = \frac{\text{Cov}(\tilde{X}^a, \tilde{Y}^a)}{\text{Var}(\tilde{X}^a)}, \tag{6}$$

where $\text{Var}(\tilde{X}^a)$ is a diagonal matrix of variances $(\sigma^2_{\tilde{x}_1} \ldots, \sigma^2_{\tilde{x}_h})$ estimated from context samples, with h the latent dimension of the sample representation x_{ia}. Note that we also apply a global translation to the bioactivity values to ensure they are centered at zero.

Given the context data and the assay representation, the model jointly predicts shape parameters (α, β) and the cluster assignment $\mathbf{w}$ through a multi-output multilayer perceptron (MLP). This is followed by soft-plus and softmax activation functions over (α, β) and $\mathbf{w}$, respectively. Here we use $K = 4$ covariance clusters. The scale parameter $\gamma = \frac{\alpha}{\beta}$ is calculated from the shape parameters of the associated gamma distribution $\text{Gamma}(\alpha, \beta)$; w is modelled with a categorical distribution.

The model input, x_{ia}, captures information on the protein sequence and ligand structure of the protein-ligand pair i within an assay a. Sequences are encoded using a convolutional neural network with a sliding window size of 20. Ligands are encoded using a graph-convolutional neural network operating on an atom-based molecular graph. The bioactivity predictions are formed by a multi-output MLP decoder, with $2K$ output channels, with the first K channels corresponding to the mean function f_k, the second K channels to the predicted variances σ^2_k. See the SI for details on the input featurization and the network architectures of the encoders and decoder.

C Training protocol

When training the model, we need to simulate an appropriate context set C using the available assay data. Chronological sampling would be ideal as this would most closely mimic the situation during hit-to-lead and lead optimization. Unfortunately, however, chronological rankings of the compounds are not usually available. Hence we resorted to on-the-fly sampling of random subsets of varying size to construct the context set, which trades in realism for improved data augmentation.

We train the model in an end-to-end fashion using the variational lower bound

$$\begin{aligned}\log p_\theta(Y_D^{1:T}|X_D^{1:T}, C) \geq \mathbb{E}_{q_\phi} \log p_\theta(Y_D^t|X_D^t, w^t, \gamma^t) \\ - D_{KL}(q_\phi(\gamma^D, w^D|D)||q_\phi(\gamma^C, w^C|C))\end{aligned} \tag{7}$$

where $D_{KL}(\cdot, \cdot)$ is the Kullback–Leibler (KL) divergence; q and p are the encoder and decoder networks, respectively. As in standard NPs, the KL divergence term encourages the network to predict scalings and weights from the context data, (γ^C, w^C) that are consistent with those from the full assay data (γ^D, w^D).

Model implementation. The model was implemented in PyTorch (33) and optimized using ADAM (34) with a learning rate of $1e^{-5}$. The graph-convolutional network was implemented using the Deep Graph Library (DGL) (35) package. RDKit (36) was used to read, write and manipulate ligands. BioPython (37) was used to read, write and manipulate protein structures and sequences. The source code and pre-trained models can be accessed at https://github.com/lucianlschan/metabind

D Baseline models.

In contrast to traditional QSAR models, MetaBind learns to aggregate relevant assays instead of relying on hard-coded point-wise aggregation rules. Furthermore, its predictions are based on local support data (the context set) and hence, by design, assay-dependent. To fairly compare MetaBind with traditional QSAR approaches, we constructed two bioactivity models, *Baseline* and *Baseline (local)* with two distinct learning protocols. Like MetaBind, both models use convolutional and graph-convolutional neural networks to encode protein sequences and ligand structures, respectively. Furthermore, the encoder architectures are identical for both the baselines and the meta-learning model. The differences in architecture are therefore on the decoder side: Specifically, in the baseline models, the protein and ligand features are concatenated and fed into an MLP to predict the ligand bioactivity (see the SI for details).

Importantly, the baselines are trained using preprocessed training data. Duplicate observations are aggregated using their geometric mean if the deviation of the minimum and maximum observed values is less than 0.3 pK units (corresponding to a threefold change in affinity) under the assumption that the endpoints are exchangeable; otherwise the observations are discarded. The two baseline models use identical network architectures, data pre-aggregation rules, and a mean-squared-error loss function for training, but differ in how they are applied to test data, as *Baseline (local)* uses a local training update over the context set to tune its weights to each test assay independently. This local version thus enables a fair comparison to MetaBind in that it produces an assay-specific SAR informed by local data.

Supplementary Information

S1 Data

Summary statistics for the assays selected from ChEMBL 30 for this study are provided in Fig. S1. The dataset includes 6968 assays, covering 830 target proteins with measured bioactivities for 100,279 molecules. The molecular weight ranges from 107 Da to 857 Da, with a median of 415 Da. The heavy-atom counts per ligand range from 8 to a maximum of 50 (as by construction). Enzymes are the predominant protein target class, followed by membrane receptors, as based on ChEMBL's classification schema. The median target sequence length is 471. The half maximal inhibition concentration, IC_{50}, and inhibition constant, K_i, are the most common endpoints, accounting for ca 63% and 31% of the data, respectively. The pK values range from -6.7 (millimolar, mM) to 2 (picomolar, pM) across all assays, with an average of -2.21.

S2 Notation and definitions

We denote by C^a and D^a, with $a = 1, ..., K$, the context and the target data from assay a, respectively, where K is the number of assays in the dataset. Further, we define $C = \cup_{a=1}^{K} C^a$ and $D = \cup_{a=1}^{K} D^a$.

The neural network components used for the convolutional and graph-convolutional neural networks include activation functions, layer normalisation, convolution and pooling operations as defined below.

Activation functions

Rectifying Linear Unit (ReLU)

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

Softplus (β = 1)

$$\text{Softplus}(x) = \frac{1}{\beta} \log\left(1 + \exp(\beta x)\right)$$

Tanh

$$\text{Tanh}(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$$

Softmax

$$\text{Softmax}(x) = \frac{\exp(x_i)}{\sum_{i=1}^{N} \exp(x_i)}$$

Convolutions

Affine transformation

$$\text{Linear}(x) = Wx + b$$

Molecular GCNN layer

$$h_i^{l+1} = \sigma\left(b^l + \sum_{j \in N(i)} \frac{1}{c_{ij}} h^{l-1} W^l\right)$$

Here, $N(i)$ is the set of neighbors of node i, c_{ij} is the square root of the product of the node degrees, σ is an activation function, l is the layer index, W^l and b^l are learnable weights and biases, respectively.

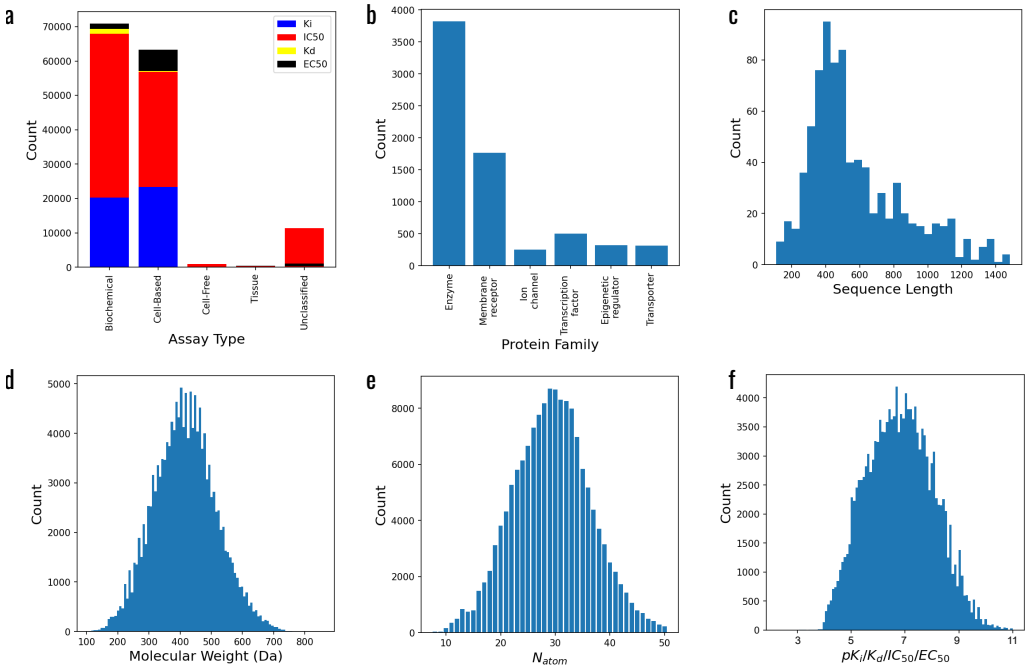

Figure S1: **Dataset summary.** (a) Composition by assay type, with biochemical and cell-based assays being the predominant types. Half maximal inhibitory concentration (IC_{50}) and inhibition constant (K_i) together make up more than 90% of the data volume. (b) Composition by protein target family. Enzymes are the predominant target class, followed by membrane receptors. (c) Composition by protein-sequence length. The median sequence length is 491. (d) Distribution of ligand molecular weights, ranging from 107Da to 857Da, with a median of 415 Da. (e) Distribution of per-ligand heavy-atom counts. The median is 29. (f) Distribution of log-affinity values. The median is 6.8.

Pooling operators

Max Pooling

$$\text{maxpool}(x^i) = \max_{k=1}^{N_i} (x_k^i)$$

Sum Pooling

$$\text{sumpool}(x^i) = \sum_{k=1}^{N_i} (x_k^i)$$

S3 Architectures

Ligand input features. Atoms (nodes) and bonds (edges) of the molecular graph are featurized using a set of chemical properties as summarized in Tab. S3 and S4. Categorical features are encoded in a one-hot format, resulting in an overall input dimension of 36 for the atom and 6 for the bond feature vectors. RDKit was used to compute all atom and bond features.

Protein input features. The residues of the protein sequence are encoded in a one-hot fashion using the amino-acid type (AA), the sidechain topology and its physicochemical properties. The four side-chain topologies are: (i) no sidechain, (ii) linear side-chain, (iii) branched side-chain, and (iv) cyclic side-chain. The physicochemical properties capture the electrostatic and hydrophobic characteristics of the different AAs: positively charge (H, K, R), negatively charged (D, E); polar (N, Q, S, T); aromatic (F, W, Y); hydrophobic (A, I, L, M, V).

Feature	Description	Dimension
Atom type	Element (one-hot)	12
Atom degree	Neighbour count (one-hot)	7
Radical electrons	Radical-electron count	1
Formal charge	Charge in e	1
Hybridization	s, sp^1-sp^3, sp^3d, sp^3d^2, other	7
Aromaticity	Aromatic indicator	1
Hydrogen count	Explicit + implicit	5
Chirality	CIP R and S indicators	2

Table S1: **Atomic input features.** Atom node attributes used by all 2D molecular embedding networks.

Feature	Description	Dimension
Bond type	Single – triple, aromatic, conjugated	5
Ring bond	Ring indicator	1

Table S2: **Bond input features.** Covalent edge attributes used by all 2D molecular embedding networks.

Ligand representation. The ligand is represented as a graph, featurized as described above, and encoded using a 3-layer graph-convolutional neural network, with residual connections between each layer. To generate the ligand features, we apply max and sum pooling over the atom encodings, and concatenate the respective outputs to obtain the ligand embedding.

Protein representation. The protein is represented as a sequence, featurized as described above, and encoded using a discrete cosine transform applied to the feature matrix, followed by a 1D convolution operation with kernel size 10 and stride 3. Note that we set a maximum sequence length of 1500 residues.

Protein-ligand representation. The joint protein-ligand embedding vector is generated from the individual ligand and protein embeddings via concatenation and subsequent linear transformation.

S4 Data splits

Paired-assay split. The paired-assay test set contains 100 assay pairs generated from 190 assays. The distribution of the number of observations per assay pair is shown in Fig. S2a. The distribution of the mean absolute deviation within a pair (i.e., the mean of the absolute ligand hererogeneity $\Delta_{ia,ib}$) is shown in Fig. S2b, and has a median of 1.05 log units. The majority of the assay pairs are positively correlated, as shown in Fig. S2c-d using the Pearson correlation r and Kendall rank correlation coefficient τ. There are multiple assay pairs, however, for which the correlation is negligible or even negative, see the tails of the distributions over r and τ.

The correlation between observed and predicted values achieved by the standard and local baseline models on the paired-assay test set is shown in Fig. S5a and b, respectively. The correlation is weak, with a Pearson r of 0.29 and 0.32, respectively. The local update on context data thus improved the model performance only slightly, limited presumably by the Lipschitz continuity/constant inherent to

Feature	Description	Dimension
Amino Acid	20 AA and 5 special characters	25
Side Chain Topology	AA side-chain topology	4
Properties	AA physicochemical properties	6

Table S3: **Protein input features.** Amino-acid (AA) features used as input for the 1D protein-sequence embedding networks.

Model	Input	Architecture	Dimensions
Ligand Encoder	L	GCN	(34,128,128)
Protein Encoder	P	CN	(32 ,32, 1)
P-L Encoder	P, L	Linear	(419,128)
Cluster Decoder	P-L,y	MLP	(128,128,6)
Bioactivity Decoder	P-L	MLP	(128,128,h)

Table S4: **Model architectures.** L, P, P-L, y denote the ligand, protein, protein-ligand and bioactivity, respectively. GCN, CN, Linear and MLP are the graph-convolutional and convolutional network, linear map and multi-layer perceptron, respectively. The values in brackets indicate the input, intermediate and output dimension, with (see bottom row) $h = 8$ for MetaBind, $h = 1$ for the baseline models.

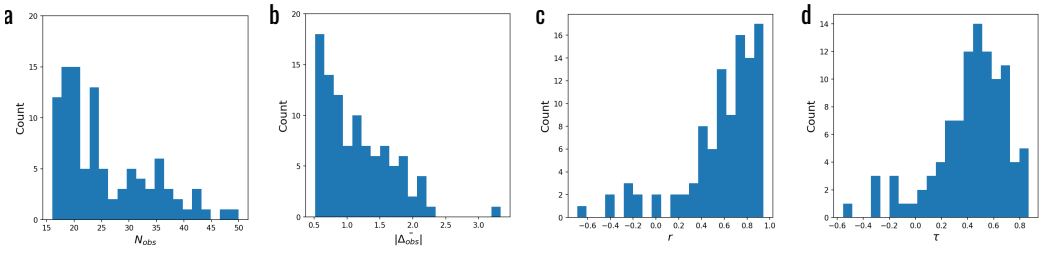

Figure S2: **Paired-assay test set.** (a) Distribution of the size of the shared ligand set (number of observations N_{obs} per assay pair. (b) Distribution of the mean absolute deviation Δ measured for the assay partners. (c) Distribution of the Pearson correlation coefficient and (d) Kendal tau of the pair read-outs. Note that some assays pairs show a negative or statistically negligible correlation.

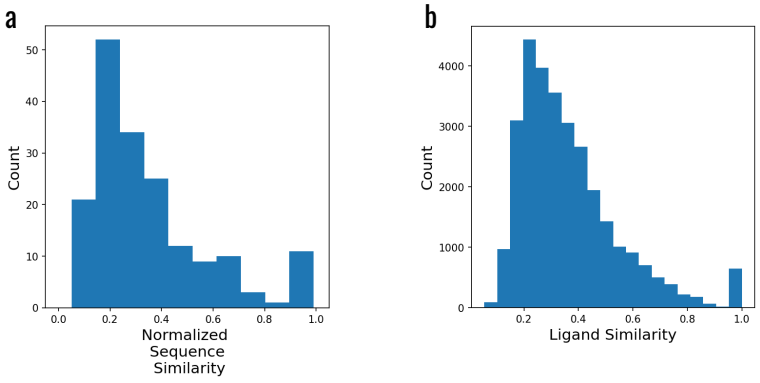

Figure S3: **Chronological test set.** (a) Distribution of the maximum sequence similarity between unobserved protein targets in the test set with protein targets of the training set. (b) Distribution of the maximum ligand similarity between ligands of the training set and ligands of the test set for all observed protein targets. The median maximum similarity is 0.32. Note that we used Tanimoto similarity over 2048-bit Morgan fingerprints of radius 2 to measure the ligand-ligand similarity.

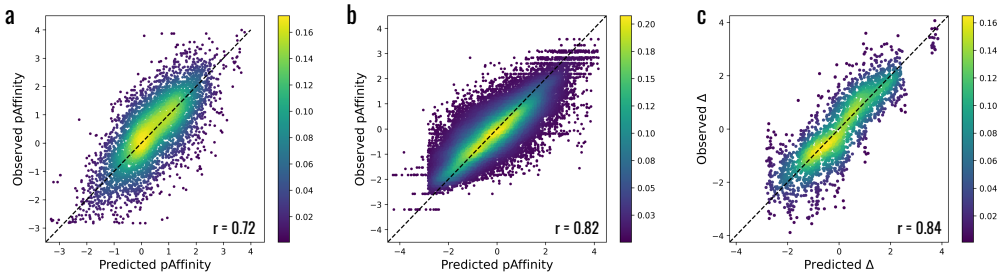

Figure S4: **MetaBind test predictions of protein-ligand affinity and per-ligand heterogeneity.** Tests based on the (a) paired-assay and (b) chronological split produce a strong Pearson correlation r of 0.72 and 0.82 with the experimental affinities, respectively. The same holds for the (c) ligand heterogeneities extracted from the test set of the paired-assay split. These heterogeneities Δ (see their definition in eq. 5) are not typically known, but here available by construction because of the congeneric series underlying the test set of the paired split. Note that panels a and b compare centred affinities to aid interpretation, with the same centring constant applied to all axes.

the neural network.

Chronological split. The test set of the chronological split contains 539 proteins, with 178 unseen targets. The distribution of the normalised sequence similarity between the unseen target proteins and observed target proteins is shown in Fig. S3a. Here the normalised sequence similarity is defined as the product of the identity score calculated from BLAST and the min-max ratio ρ of the sequence lengths,

$$\rho_{AB} = \frac{\min(N_{AA}^{A}, N_{AA}^{B})}{\max(N_{AA}^{A}, N_{AA}^{B})},$$

where N_{AA}^{A}, N_{AA}^{B} are the lengths of sequences A and B, respectively. We used default parameters for the BLAST. For assays with targets already seen during training, we calculated the maximum ligand similarity between the training set and test set of those targets, with the resulting distribution of similarities shown in Fig. S3b. Even though some assays share identical ligands, most of the ligand sets differ considerably, as indicated by a relatively low median ligand similarity of 0.32.

As for the paired-assay test set, both the standard and local baseline give poor fits with a limited dynamic range and Pearson r of 0.36 and 0.39, respectively (see Fig. S5c-d).

Assay	ChEMBL ID	Description
A	CHEMBL1039372	Inhibition of human recombinant renin assessed as decrease in plasma renin activity by competitive radioimmunoassay in presence of human plasma
B	CHEMBL1039373	Inhibition of trypsin-activated human recombinant renin
C	CHEMBL2073500	Inhibition of human KDR autophosphorylation expressed in mouse NIH/3T3 cells
D	CHEMBL2073498	Inhibition of KDR by HTRF analysis in presence of 1 mM ATP
E	CHEMBL3804128	Inhibition of wild type B-Raf in human MIAPaCa2 cells assessed as reduction in ERK phosphorylation preincubated for 1 hr by Western blot method
F	CHEMBL3804127	Inhibition of B-Raf V600E mutant (unknown origin) assessed as MEK1 phosphorylation using MEK1-Avitag as substrate after 1 hr by HTRF assay

Table S5: ChEMBL ID and assay description of assays A-F used to illustrate the different heterogeneity classes associated with the pairs (A,B) – constant offset; (C,D) – uncorrelated; (E,F) – offset and rescaling.

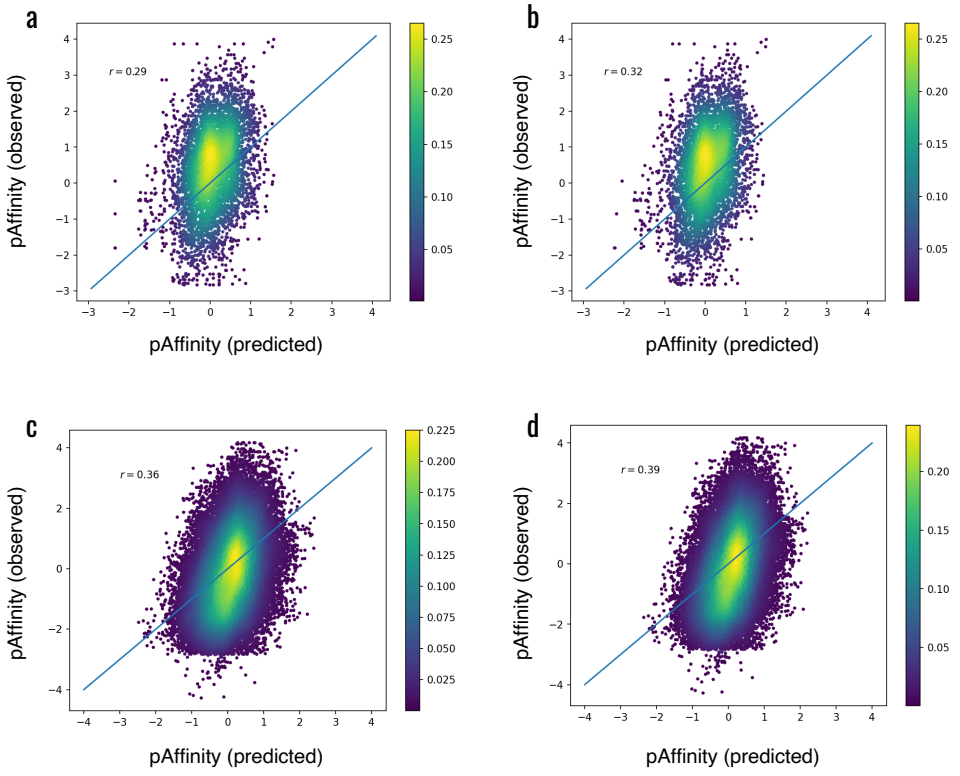

Figure S5: **Baseline test predictions of protein-ligand affinity.** Correlation plots comparing observed and predicted affinity values for the (a) standard baseline and (b) local baseline under the paired-assay split; (c) the standard baseline and (d) local baseline under the chronological split. To aid interpretation, all panels compare centred affinities, with the same centring constant applied to all axes.