Position: AI for Drug Discovery Models Often Do Not Learn as Expected and How to Diagnose These Failure Modes

Anonymous Authors[1]

Abstract

We argue that in multiple areas of AI for drug discovery (AIDD), deep learning models are not learning the meaningful biological or chemical features they were hypothesised to capture, but instead learn non-generalisable features, for example, from dataset biases. To address this, we propose the systematic use of *misaligned baselines*. We define misaligned baselines as models that rely on signals purposely misaligned with the intended learning objective. Rather than modelling the underlying biology or chemistry directly, these baselines draw on reductionist sources of information, heavily perturbed input features, or heuristic models. Competitive performance of such baselines reveals when models are not learning the biologically meaningful representations they were designed to learn. By examining case studies across multiple branches of AIDD, we demonstrate that misaligned baselines have consistently exposed such failure modes and crucially informed the evaluation and improvement of these models. We argue that adopting them as standard practice will help to ensure that progress in AIDD reflects genuine advances rather than artefacts of data-generating processes or evaluation design.

1. Introduction

AI for drug discovery (AIDD) has great promise to accelerate the development of novel therapeutics while reducing costs and decades-long timelines that characterise traditional drug development (Paul et al., 2010; Walters & Barzilay, 2021; Didi et al., 2026; Branson & Deane, 2025; Dreyer et al., 2025; Mille-Fragoso et al., 2025). Deep learning models have been increasingly deployed across critical stages of the drug discovery pipeline. However, despite widespread adoption and reported state-of-the-art performance across numerous tasks, a growing body of evidence suggests that many of these models often fail to learn the meaningful biological and chemical features and relationships they were designed to capture (Branson et al., 2025; Wenteler et al., 2025; Durant et al., 2025; Ahlmann-Eltze et al., 2025; Janela & Bajorath, 2022; Svatko et al., 2026).

The central problem this paper aims to address is *performance misalignment*, which refers to the disconnect that exists between what models are actually learning and what researchers have assumed they are learning based on model performance. The existence of such misalignment not only limits advances in AIDD but also leads to the usage of these models with unrealistic performance expectations. To address this critical issue, we propose the systematic use of *misaligned baselines*.

Misaligned baselines definition. We broadly define misaligned baselines to be baselines that are fundamentally misaligned with the intended learning objective of the models they are compared to. Specifically, we consider simple heuristic-based baselines, such as predicting the mean gene expression profile for a given perturbation, as well as feature and model-constrained baselines that deliberately limit the model's access to potentially informative representations from the input features or model, respectively. For instance, baselines that utilise only low-dimensional molecular descriptors (e.g., 1D/2D chemical fingerprints) for tasks where sophisticated deep learning models were hypothesised to extract critical structural information from high-dimensional molecular representations or 3D conformational features. Thus, rather than modelling the underlying biological or chemical mechanisms that deep learning models were originally designed to capture, these misaligned baselines exploit alternative sources of signal such as target value distributions, dataset artefacts, or predictive signal extracted from deliberately constrained input features or model representations.

Misaligned baselines serve three critical functions that make them well-suited for AIDD evaluation:

1. **Reveal when models are not learning what is expected:** When sophisticated models perform compa-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

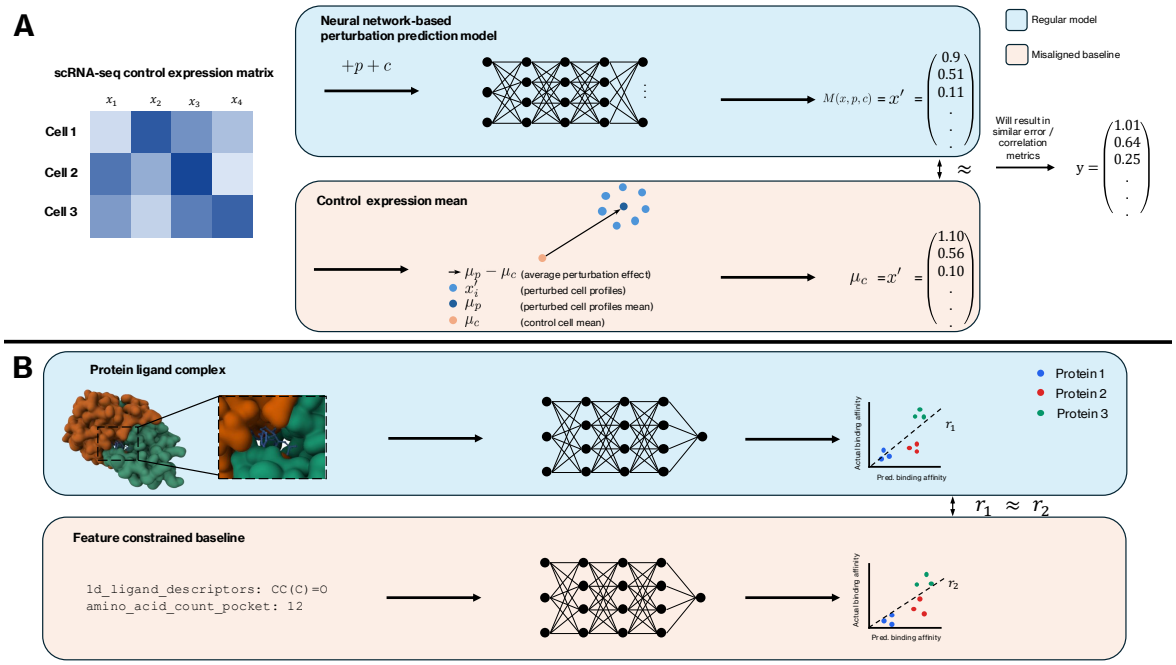

Figure 1. **Examples of misaligned baselines (A)** Transcriptomic perturbation prediction. A deep model predicts the perturbed profile x' from a control profile x, perturbation p, and covariates c. The control expression mean baseline simply predicts $\hat{x}' = \mu_c$. When the average perturbation effect $\mu_p - \mu_c$ is small, which is particularly true in high-dimensional evaluation settings, both give similar predictions (see section 3.1). **(B)** Structure-based binding affinity prediction. A deep model trained on 3D protein–ligand poses achieves correlation r_1. A feature-constrained baseline using only 1D ligand descriptors and pocket residue counts achieves $r_2 \approx r_1$ (see section 3.3).

rably to, or worse than, misaligned baselines, this exposes a fundamental disconnect between the hypothesised learning mechanism and the actual learning mechanism.

2. **Providing essential performance context:** Misaligned baselines ground performance metrics in meaningful terms by establishing what level of performance can be achieved without learning the intended relationships. This context is crucial because raw performance metrics in isolation often fail to reflect true model quality. For instance, a model achieving $R^2 = 0.85$ in predicting perturbation effects appears highly successful until a mean-prediction baseline achieves a similar R^2 by exploiting the fact that most genes show minimal expression changes. This grounding can also help with designing better performance metrics.

3. **Diagnosis and resolution of failure modes:** Misaligned baselines can be used as evidence that the model may be exploiting dataset artefacts, memorising target distributions, or learning spurious correlations rather than the intended biological relationships. This allows researchers to develop methods to resolve these issues and improve their models.

We argue that the adoption of misaligned baselines as standard practice is vital for progress in AIDD. By exposing performance misalignment in the models, these baselines enable researchers to identify and prioritise the most valuable avenues for model improvement, ensure realistic expectations when deploying systems in real-world applications, and prevent wasted resources on development efforts guided by deceptively strong metrics that mask fundamental learning failures.

2. Background

In this section, we provide a background of the prediction areas that we cover in this paper: transcriptomic perturbation response prediction, drug response prediction, binding affinity prediction and transfer learning from protein language models. These areas are used as examples of the importance of continued inclusion of such baselines. For more detailed descriptions of any of these areas, we refer readers to the cited works in the respective subsections.

2.1. Transcriptomic perturbation prediction

Transcriptomic perturbation prediction represents a fundamental challenge in computational biology: predicting how a cellular transcriptome changes in response to a perturbation. Reliably forecasting transcriptomic perturbation

effects is an essential cornerstone for building a virtual cell, an *in silico* model for simulating cellular behaviour. The ability to accurately predict perturbation responses holds immense promise for drug discovery, personalised medicine, and understanding cellular mechanisms without having to run expensive and time-consuming experimental screens (Bunne et al., 2024).

In this section, a *perturbation* encompasses any intervention that leads to transcriptomic alteration of a cell, including genetic manipulations (e.g., CRISPR-mediated gene knockout or overexpression), chemical treatments (e.g., small-molecule drugs), cellular differentiation cues (e.g., TGF-β signalling) or environmental changes (e.g., osmotic stress).

Problem formulation. The core prediction task can be formalised as learning a mapping from an unperturbed cellular state to its corresponding perturbed state. Let x represent the gene expression profile of a control cell. Given a perturbation p and covariates c the objective is to predict the resulting perturbed expression profile x' We can express this model as:

$$x' = M(x, p, c) \tag{1}$$

Typically, models used for transcriptomic perturbation prediction include linear models (Kamimoto et al., 2023; Gaudelet et al., 2024), graph neural networks that incorporate gene regulatory network structure (Roohani et al., 2023), and foundation models, with transformer-type architectures, pre-trained on large-scale single-cell datasets (scFMs) (Cui et al., 2024; Hao et al., 2024; Theodoris et al., 2023; Adduri et al., 2025; Gong et al., 2023). scFMs aim to capture complex, nonlinear relationships between perturbations and cellular responses through learned representations of both genes and perturbations and have reported state-of-the-art performance.

2.2. Cellular drug response prediction

Being able to accurately predict the efficacy of anti-cancer drugs promises to help speed up the drug discovery process and provide a vitally needed stratified approach to cancer treatments (Toniatti et al., 2014). Thus, the aim of anti-cancer drug response prediction (DRP) is to predict the efficacy metrics for small molecules against different cancer subtypes, where the subtypes are described using their omics profiles (Partin et al., 2021).

Problem formulation. Typically, DRP is done in the context of cancer cell lines and $IC50$ values are predicted based on the omics profiles of the cell lines (Sharifi-Noghabi et al., 2021). Focusing on cancer cell lines, we define DRP as building a model M that is able to accurately make predictions of drug efficacy for the cell line i and drug j, such that

$$\hat{y}_{i,j} = M(x_j^d, x_i^c). \tag{2}$$

Where $\hat{y}_{i,j}$ is the model's prediction of $y_{i,j}$ the efficacy of drug j for cancer cell line i. The model takes as input x_j^d, the representation of drug j and x_i^c, the representation of cell line i. Example inputs are smiles strings and transcriptomics for the drug and cell line profiles, respectively. In this paper, we focus on deep learning models, as they have reported state-of-the-art performance in DRP (Baptista et al., 2021). There is a wide array of DL architectures used, including pre-trained transformers, graph neural networks, and convolutional neural networks (Branson et al., 2023; 2025; Partin et al., 2023; Hostallero et al., 2022).

There are three common applications of DRP models corresponding to three different ways of training and testing the models and possible because there are both drug and cell line inputs to the model.

1. **Mixed set.** All input drugs and cell lines are in both training and testing sets but unique drug cell line pairs are only in one.
2. **Cancer blind.** A subset of cell lines are held out during training and tested on.
3. **Drug blind.** A subset of drugs are held out during training and tested on.

Mixed set tests a model's ability to repurpose existing drugs for known cancer subtypes. Cancer blind tests if the model could be used to speed up drug screening and to rank drugs for unseen subtypes. Drug blind evaluates if a model can predict the efficacy of a novel drug just from its structure.

2.3. Binding affinity prediction between small molecule drugs and proteins

2.3.1. Structure-dependent scoring functions

Binding affinity is a key property for a lead compound in early-stage drug discovery and is crucial to maintain while other properties are also optimised, such as absorption, distribution, metabolism, excretion and toxicity (ADMET) (Volkov et al., 2022). Machine learning has been applied in different ways to predict this property for a given target protein with varying success. Machine learning-based scoring functions (MLBSF) are a common method for binding affinity prediction (Durant et al., 2025). These MLBSFs take the structural information of a bound protein ligand complex to predict binding affinity. MLBSFs learn the relationship between static protein-ligand complex structure and its binding affinity. They learn this from datasets with experimentally determined crystal structures and curated affinity

measurements, such as PDBBind (Wang et al., 2005). The hope is that these methods are more generalisable as they have learnt concepts of physics from the static structures that are transferable to any novel protein family or ligand chemical space.

Problem formulation. This problem can be formulated simply as:

$$\hat{y} = M(x_p, x_l) \tag{3}$$

where x_p and x_l denote the 3D structures of the protein and ligand, respectively, in their bound conformation. These structures should not be independent, as they reflect mutual spatial and chemical adaptation within the complex. However, often the protein conformation is independent of the ligand if docked into the rigid protein. The model M is trained to map the joint structural context to a predicted binding affinity $\hat{y}$.

2.3.2. Machine learning QSAR models

Another, common approach is to build models that predict binding affinity for a single protein, and so the model is only given information about the 2D representation of the ligand (Walters & Barzilay, 2020). These types of models have been built using older methods like statistical models to create Quantitative Structure-Activity Relationship (QSAR) models. Newer machine learning models aim to learn richer and more expressive representations of the ligand to predict binding affinity more accurately.

Problem formulation. Using similar notation to above:

$$\hat{y} = M_p(y_l) \tag{4}$$

where y_l is the 2D representation of a ligand, and M_p is a model specific to a given protein target p. The model predicts $\hat{y}$, the binding affinity between ligand y_l and the fixed protein p.

2.4. Transfer learning from protein language models

Protein language models (PLMs) are typically transformer-based models trained with masked language modelling on large amounts of unlabelled amino acid sequences of proteins (Hayes et al., 2025; Xiao et al., 2025; Turnbull et al., 2024). These pre-trained models can then be fine-tuned on relatively small amounts of labelled data for specific tasks.

Problem formulation. Given a pre-trained PLM that has then been fine-tuned on task-specific data $\{x_s, y\}$, M, we want to predict:

$$\hat{y} = M(x_s)$$

,

where x_s is the amino acid sequence of a protein and y is its associated label. For example, y could be the thermostability of the protein x, a key property to account for when designing antibody therapeutics (Rollins et al., 2024). It is hypothesised that features learned in pre-training will transfer to a given downstream task, improving performance.

3. Findings from misaligned baselines

This section describes case studies in each of the areas above, where misaligned baselines have revealed that deep learning models were not learning as expected. We organise each case study using a consistent framework. First, we describe the initial hypothesis about what features the deep learning models were expected to learn; second, we explain how a misaligned baseline could theoretically solve the same task without learning these intended features; finally, we present the empirical findings that demonstrate the misaligned baselines can rival or exceed the deep learning models' performance.

3.1. Transcriptomic perturbation prediction

In this section, we detail various concurrent findings on the performance of transcriptomic perturbation prediction models We focus specifically on the task, of training models to predict unseen perturbations.

Originally hypothesised learning mechanism. By training on control–treatment pairs of gene expression, perturbation identity, and optionally covariates, the model learns a biologically meaningful representation for predicting unseen perturbations (Bunne et al., 2024; Roohani et al., 2023; Cui et al., 2024).

Misaligned baseline hypothesis. scRNA-seq is characterised by its high dimensionality, technical and biological noise, and sparsity. Furthermore, many perturbations have a concentrated effect on the transcriptome, meaning that most gene expression values remain relatively unchanged post-perturbation. Therefore, using the average expression across the control cells is already a strong predictor of transcriptomic perturbation response.

Misaligned baseline results. The average baseline performs comparably to foundation models and perturbation-specific models that report SOTA performance for unseen single gene perturbation prediction (Ahlmann-Eltze et al., 2025). Similarly, in (Wenteler et al., 2025), the authors found that the zero-shot embeddings from a wide range of single-cell foundation models did not improve performance over the mean baseline. These findings stem from mode collapse driven by the high dimensionality and sparsity of RNA-seq data, combined with typically small effect sizes

from perturbations. Under such conditions, models are incentivised to approximate the average perturbation response since the majority of gene expression values remain unchanged. Simple misaligned baselines, such as predicting no change (null baseline) or the average control expression (mean baseline), therefore perform competitively with complex deep learning architectures, a finding that has been confirmed in numerous benchmark studies across various model architectures (Ahlmann-Eltze et al., 2025; Wenteler et al., 2025; Wu et al., 2025; Kernfeld et al., 2024; Csendes et al., 2024).

Following these findings, the community rapidly adopted these baselines in benchmarking efforts and developed more biologically meaningful evaluation metrics to address these limitations. Adduri et al. (Adduri et al., 2025) introduced *Cell-Eval*, an evaluation framework incorporating biologically meaningful metrics alongside misaligned baselines such as *PerturbMean* and *ContextMean*. Another example of the progress inspired by failure modes revealed by misaligned baselines is by Mejia et al. (Mejia et al., 2025), where the authors directly addressed the mode collapse problem by investigating evaluation metrics that better capture biological diversity in perturbation responses, moving beyond simple correlation-based measures that reward averaging behaviours. In (Viñas Torné et al., 2025), Viñas et al. propose redefining the evaluation reference of perturbation effect from control cell expression to the global perturbation mean, effectively neutralising the performance of the baseline that uses the average expression across control cells.

3.2. Cellular drug response prediction

In (Branson et al., 2025), Branson et al. used different misaligned baselines for the different testing types discussed in section 2.2, and compared them to deep learning-based models that had reported SOTA performance for DRP. We detail the findings for the testing types below.

3.2.1. Cancer blind testing

Originally hypothesised learning mechanism. Features predictive of drug efficacy are learnt from the input omics profiles, and chemical drug structures.

Misaligned baseline hypothesis. There are generally drugs that are effective or ineffective for a range of cancer subtypes. Thus, the average efficacy of each of the drugs in the training set can capture this, leading to strong performance, without using cell line profiles or chemical drug structures.

Misaligned baseline results. The baselines can outperform the DL models that use genomics cell line profiles. Thus, showing these models are not learning predictive features from the input data as was expected. The authors also found that models using transcriptomics features were able to outperform the baseline, but that much of the performance of these models can be explained by their learning the average efficacy of the drugs.

Branson et al. (Branson et al., 2025) also found that by testing across cell lines for a given drug, which is not currently commonplace in DRP, the misaligned baseline gave the performance naively expected by such a baseline ($R^2 \sim 0$), and in this setting, some of the models had a much larger performance gap w.r.t this baseline. Thus, suggesting DRP models can have the most impact in this testing regime, opening up an exciting avenue of research, partly motivated by a misaligned baseline. Additionally, in (Branson et al., 2023), the authors demonstrated that as dataset size increases, the performance of the average baseline plateaus, but models trained using transcriptomics and proteomics data did not. This suggests the performance differential to the baseline can be widened by collecting more data

3.2.2. Mixed set testing

For this testing type, originally hypothesised learning mechanisms are the same as above.

Misaligned baseline hypothesis. As well as having generally effective drugs, there are CLs that are generally susceptible or resistant across drugs. Thus, because for mix set testing, all CLs and drugs are in both the train and test sets (unique drug-CL pairs are only in one set), a model can be trained to directly learn this from the training ground truth distribution. To achieve this, the authors defined and created the *marker model* by replacing the chemical and biological input data with marker inputs, representations defined to have no biological or chemical information e.g., with unique one-hot encoded column vectors representing each input.

Misaligned baseline results. The marker model outperformed the DL models, which had reported SOTA performance for DRP. Thus, rather than learning biological or chemically relevant features, the marker model learned a function that simply leveraged the training target values to great success. Again, here the baseline suggests that the models are not learning what was expected (Jiang et al., 2022; Liu et al., 2019; Nguyen et al., 2022; Baptista et al., 2021) and instead of learning biological and chemically relevant features, they are leveraging the distribution of the target values.

The authors hypothesised that some of these limitations were due to the noisy nature of the target values (Haibe-Kains et al., 2013; Safikhani et al., 2017), which meant predictive features from the input data could not be learned. By binarising the response values, they then showed that for mixed set testing, the models were able to slightly outperform the baseline, suggesting a key limitation was the data quality. This shows that misaligned baselines can be

used to identify failure modes that can, in turn, be mitigated, leading to improved models.

3.3. Binding affinity prediction between small molecule drugs and proteins

3.3.1. Structure-dependent scoring functions

Originally hypothesised learning mechanism. Features predictive of binding affinity are learnt from the 3D bound pose of the protein and ligand (Ballester & Mitchell, 2010; Stepniewska-Dziubinska et al., 2018; Scantlebury et al., 2023; Li et al., 2021; Wang et al., 2021)

Misaligned baseline hypothesis. In datasets used for structure-based binding affinity prediction, there are ligands that have a similar binding propensity across multiple proteins, and proteins that are generally easier or harder to bind to for multiple ligands. Thus, these characteristics can be learnt directly from the identities of protein and ligand without the models learning any predictive information specifically from the 3D poses. To test this in (Durant et al., 2025), Durant et al. use simple 1D/2D descriptors of the ligands and amino acid counts of the protein pocket as inputs to their baseline instead of the 3D inputs.

Misaligned baseline results. The baseline can match or outperform deep learning based models that have reported SOTA performance for structure-based binding affinity prediction (Durant et al., 2025). This suggests that the models are not learning representations grounded in biophysics, as hypothesised initially (Ballester & Mitchell, 2010; Stepniewska-Dziubinska et al., 2018; Scantlebury et al., 2023; Li et al., 2021; Wang et al., 2021), but instead exploiting more simplistic heuristics. This finding is also supported by a separate study (Volkov et al., 2022).

More recently, success has been reported in accurately predicting binding affinity by training on large datasets curated from databases such as BindingNet and ChEMBL (Valsson et al., 2025; Passaro et al., 2025), showing that increasing dataset size is key to learning these more complex representations.

3.3.2. Machine learning QSAR models for binding affinity

Originally hypothesised learning mechanism. The models are learning rich and expressive representations of the ligand in order to predict binding affinity (pIC_{50}) (Tropsha et al., 2024).

Misaligned baseline hypothesis. Similarly structured ligands show similar properties. Thus, for a given ligand, simply using the target values of the k most similar ligands from the training set leads to similar performance as the more complex ML and DL based models. Here, Janela et al. (Janela & Bajorath, 2022) simply used the target values of the k most similar ligands from the training set, using a k-nearest neighbours baseline.

Misaligned baseline results. The k-nearest neighbours baseline outperforms graph convolutional networks and other ML-based models (Janela & Bajorath, 2022). This shows that the models were not learning performative features from the input ligand that exceed what can be found from the target values of the most similar drugs in the training set.

This baseline highlighted the need for activity cliff data, which consist of small changes in ligand structure but dramatic changes in binding affinity, in the training data. Learning from this data would adversarially penalise models that resort to learning similarity and so learn more complex relationships. The field is moving to adopt learning from this type of data, and in doing so is improving model accuracy (Ibragimova et al., 2025; Gu et al., 2025).

3.4. Transfer learning of protein language models

Originally hypothesised learning mechanism. Protein language models learn features during unsupervised pre-training that help improve performance on downstream tasks (Rives et al., 2021; Wang et al., 2024). One class of problems PLMs are applied to is variant effect prediction. Where properties such as binding affinity, thermostability and expression of different variants of a given protein. These properties can be referred to as the protein fitness. Here, we focus on the task of predicting the fitness of different variants of the adeno-associated virus capsid protein, as in (Li et al., 2024).

Misaligned baseline hypothesis. Model performance is due to fine-tuning on the specific task. Thus, starting from a model with randomly initialised weights would give the same performance as fine-tuning starting from a pre-trained model.

Misaligned baseline results. For the adeno-associated virus task, after fine-tuning both the pre-trained PLMs and ones with randomly initialised weights, Li et al. (Li et al., 2024) found that the pre-trained PLM failed to consistently outperform the baseline. Thus, showing the models were not learning reliably transferable features in pre-training. Li et al. further found that out of the other tasks they considered (e.g. thermostability, binding and structure prediction), structure prediction was the only task where the performance of the models improved as the pre-trained model size increased. These results give both vital context on how much performance is due to the pre-training and isolate the failure mode to a combination of the adeno-associated virus dataset and the PLM rather than a fundamental issue with PLM transfer learning.

4. Discussion

General findings from misaligned baselines. In the context of AIDD, the core question that the misaligned baselines test is whether the models are learning meaningful biological or chemical features, or if the performance is due to another factor. We expect and want our models to learn some of the factors captured by misaligned baselines. For instance, the biological reality that most perturbations have small effects should inform model predictions. The problem arises when this constitutes *all* that the model is learning, rendering sophisticated architectures no more informative than simple heuristics and unable to capture the nuanced biological relationships we would like our models to capture. As we describe in this paper, across multiple very different areas of drug discovery, we frequently observed that deep learning models perform comparably, or even worse, than deliberately simplistic baselines. This pattern reveals a fundamental disconnect between the hypothesised learning mechanisms and the actual mechanisms these models employ to achieve their reported performance.

The implications extend beyond individual model failures. These findings suggest that many reported advances in AIDD may reflect improvements in exploiting dataset artefacts rather than genuine progress in learning biologically meaningful representations. Fortunately, the use of misaligned baselines and the failings they identify are already having a positive impact. For example, in transcriptomic perturbation prediction, where misaligned baselines have spurred innovation in modelling and evaluation strategies that have already reduced performance misalignment.

Recommendations for creating misaligned baselines. There are areas of study in AIDD and beyond where misaligned baselines have not yet been employed, but which would greatly benefit from misaligned baselines. Due to the unique aspects of each area, there is not a one-size-fits-all misaligned baseline that can be applied across all of them. Thus, here we recommend some directions to explore when developing a misaligned baseline for a problem, or applying a misaligned baseline in a new context.

1. **Domain specialist heuristics:** What would a specialist in the field come up with as a rule-of-thumb prediction given the same problem and data? A misaligned baseline derived from this can contextualise model performance. As demonstrated in our case studies, such baselines can already capture a substantial portion, or even all, of model performance, indicating that complex models may be learning little beyond what simple heuristics already elucidate.

2. **Feature perturbation:** What perturbations to the input features can be made to test whether the model is learning as hypothesised? For example, by constraining the input data, such that there's no 3D information, for models that are designed to learn to exploit this structural information (Section 3.3). At the extreme, this can be replacing biologically and chemically relevant features with unique identifier vectors that simply mark each input, as demonstrated in the marker model approach (Section 3.2.2). This tests whether models are learning meaningfully from the input features or merely memorising training set characteristics.

5. Conclusion

In this paper, we have argued that many AIDD models are not learning as expected, and that *misaligned baselines*, models that are misaligned with the intended learning objective, can and should be used to uncover this performance misalignment and push the field forward. We have demonstrated the utility of these baselines across transcriptomic perturbation prediction, drug response prediction, binding affinity prediction, and PLM transfer learning, showing that they consistently expose fundamental learning failures in state-of-the-art models and provide essential context for interpreting performance. By systematically adopting misaligned baselines as standard practice, the community could: (1) identify when sophisticated models are not learning beyond simple heuristics, (2) provide meaningful performance context that prevents overinterpretation of raw metrics, and (3) diagnose specific failure modes to guide targeted improvements. We recommend that misaligned baselines be adopted as standard practice in AIDD evaluation. This approach would help researchers to diagnose and subsequently fix when their models are not learning as expected, as well as help ensure that reported advances reflect genuine progress in learning biologically and chemically meaningful representations.

Beyond AIDD, these principles can apply to any domain where models are expected to learn complex, domain-specific relationships from high-dimensional data with inherent structure and biases. Areas such as climate modelling and materials science, where models must distinguish genuine scientific relationships from spurious correlations, would likely benefit from similar baseline methodologies. The key is developing baselines that test the specific learning assumptions underlying each field's modelling approaches.

References

Adduri, A. K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., Ricci-Tam, C., Carpenter, C., Subramanyam, V., Winters, A., Tirukkovular, S., Sullivan, J., Plosky, B. S., Eraslan, B., Youngblut, N. D., Leskovec, J., Gilbert, L. A., Konermann, S., Hsu, P. D., Dobin, A., Burke, D. P., Goodarzi, H., and Roohani, Y. H. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. URL https://www.biorxiv.org/content/early/2025/07/10/2025.06.26.661135.

Ahlmann-Eltze, C., Huber, W., and Anders, S. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines. *bioRxiv*, 2025. doi: 10.1101/2024.09.16.613342. URL https://www.biorxiv.org/content/early/2025/02/07/2024.09.16.613342.

Ballester, P. J. and Mitchell, J. B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

Baptista, D., Ferreira, P. G., and Rocha, M. Deep learning for drug response prediction in cancer. *Briefings in bioinformatics*, 22(1):360–379, 2021.

Branson, N. and Deane, C. Antidif: Accurate and diverse antibody specific inverse folding with discrete diffusion. *bioRxiv*, pp. 2025–07, 2025.

Branson, N., Cutillas, P. R., and Bessant, C. Comparison of multiple modalities for drug response prediction with learning curves using neural networks and xgboost. *Bioinformatics Advances*, pp. vbad190, 2023.

Branson, N., Cutillas, P. R., and Bessant, C. Understanding the sources of performance in deep drug response models reveals insights and improvements. *Bioinformatics*, 41 (Supplement_1):i142–i149, 2025.

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., Al-Quraishi, M., Brennan, P., Burkhardt, D. B., Califano, A., Cool, J., Dernburg, A. F., Ewing, K., Fox, E. B., Haury, M., Herr, A. E., Horvitz, E., Hsu, P. D., and Jain, V. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, Dec 2024. doi: https://doi.org/10.1016/j.cell.2024.11.015. URL https://www.cell.com/cell/fulltext/S0092-8674(24)01332-1.

Csendes, G., Szalay, K. Z., and Szalai, B. Benchmarking a foundational cell model for post-perturbation RNAseq prediction. *bioRxiv*, pp. 2024.09.30.615843, January 2024. doi: 10.1101/2024.09.30.615843. URL http://biorxiv.org/content/early/2024/10/01/2024.09.30.615843.abstract.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, Feb 2024. doi: https://doi.org/10.1038/s41592-024-02201-0.

Didi, K., Zhang, Z., Zhou, G., Reidenbach, D., Cao, Z., Cha, S., Geffner, T., Dallago, C., Tang, J., Bronstein, M. M., et al. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dreyer, F. A., Schneider, C., Kovaltsuk, A., Cutting, D., Byrne, M. J., Nissley, D. A., Kenlay, H., Marks, C., Errington, D., Gildea, R. J., et al. Computational design of therapeutic antibodies with improved developability: efficient traversal of binder landscapes and rescue of escape mutations. In *MAbs*, volume 17, pp. 2511220. Taylor & Francis, 2025.

Durant, G., Boyles, F., Birchall, K., Marsden, B., and Deane, C. M. Robustly interrogating machine learning-based scoring functions: what are they learning? *Bioinformatics*, 41(2):btaf040, 2025.

Gaudelet, T., Vecchio, D., Carrami, E. M., Cudini, J., Kapourani, C.-A., Uhler, C., and Edwards, L. Season combinatorial intervention predictions with Salt & Peper. *arXiv*, 2024. URL https://arxiv.org/abs/2404.16907.

Gong, J., Hao, M., Cheng, X., Zeng, X., Liu, C., Ma, J., Zhang, X., Wang, T., and Song, L. xtrimogene: an efficient and scalable representation learner for single-cell rna-seq data. *Advances in Neural Information Processing Systems*, 36:69391–69403, 2023.

Gu, Y., Xia, S., Ouyang, Q., and Zhang, Y. Complex structure-free compound-protein interaction prediction for mitigating activity cliff-induced discrepancies and integrated bioactivity learning. *ChemRxiv*, 2025.

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., and Quackenbush, J. Inconsistency in large pharmacogenomic studies. *Nature*, 504 (7480):389–393, 2013.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):

1481–1491, Jun 2024. doi: https://doi.org/10.1038/s41592-024-02305-7. URL https://www.nature.com/articles/s41592-024-02305-7.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Hostallero, D. E., Li, Y., and Emad, A. Looking at the big picture: incorporating bipartite graphs in drug response prediction. *Bioinformatics*, 38(14):3609–3620, 2022.

Ibragimova, R., Iliadis, D., and Waegeman, W. Enhancing drug-target interaction prediction through transfer learning from activity cliff prediction tasks. *Journal of Chemical Information and Modeling*, 65(13):6558–6567, 2025.

Janela, T. and Bajorath, J. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. *Nature Machine Intelligence*, 4(12):1246–1255, 2022.

Jiang, L., Jiang, C., Yu, X., Fu, R., Jin, S., and Liu, X. Deeptta: a transformer-based model for predicting cancer drug response. *Briefings in bioinformatics*, 23(3): bbac100, 2022.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, Feb 2023. doi: https://doi.org/10.1038/s41586-022-05688-9.

Kernfeld, E., Yang, Y., Weinstock, J. S., Battle, A., and Cahan, P. A systematic comparison of computational methods for expression forecasting. *bioRxiv*, 2024. doi: 10.1101/2023.07.28.551039. URL https://www.biorxiv.org/content/early/2024/10/01/2023.07.28.551039. Publisher: Cold Spring Harbor Laboratory.

Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K., and Lu, A. X. Feature reuse and scaling: Understanding transfer learning with protein language models. In *Forty-first International Conference on Machine Learning*, 2024.

Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., and Xiong, H. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 975–985, 2021.

Liu, P., Li, H., Li, S., and Leung, K.-S. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC bioinformatics*, 20(1): 1–14, 2019.

Mejia, G. M., Miller, H. E., Leblanc, F. J. A., Wang, B., Swain, B., and de Lima Camillo, L. P. Diversity by Design: Addressing Mode Collapse Improves scRNA-seq Perturbation Modeling on Well-Calibrated Metrics. *arXiv*, 2025. URL https://arxiv.org/abs/2506.22641.

Mille-Fragoso, L. S., Wang, J. N., Driscoll, C. L., Dai, H., Widatalla, T., Zhang, X., Hie, B. L., and Gao, X. J. Efficient generation of epitope-targeted de novo antibodies with germinal. *bioRxiv*, 2025.

Nguyen, T., Nguyen, G. T., Nguyen, T., and Le, D. H. Graph Convolutional Networks for Drug Response Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):146–154, 2022. ISSN 15579964. doi: 10.1109/TCBB.2021.3060430.

Partin, A., Brettin, T., Evrard, Y. A., Zhu, Y., Yoo, H., Xia, F., Jiang, S., Clyde, A., Shukla, M., Fonstein, M., et al. Learning curves for drug response prediction in cancer cell lines. *BMC bioinformatics*, 22:1–18, 2021.

Partin, A., Brettin, T. S., Zhu, Y., Narykov, O., Clyde, A., Overbeek, J., and Stevens, R. L. Deep learning methods for drug response prediction in cancer: predominant and emerging trends. *Frontiers in Medicine*, 10:1086097, 2023.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pp. 2025–06, 2025.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Rollins, Z. A., Widatalla, T., Waight, A., Cheng, A. C., and Metwally, E. Ablef: antibody language ensemble fusion for thermodynamically empowered property predictions. *Bioinformatics*, 40(5):btae268, 2024.

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, pp. 1–9, Aug 2023. doi: https://doi.org/10.1038/s41587-023-01905-6. URL https://www.nature.com/articles/s41587-023-01905-6.

Safikhani, Z., Smirnov, P., Freeman, M., El-Hachem, N., She, A., Rene, Q., Goldenberg, A., Birkbak, N. J., Hatzis, C., Shi, L., et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Research*, 5:2333, 2017.

Scantlebury, J., Vost, L., Carbery, A., Hadfield, T. E., Turnbull, O. M., Brown, N., Chenthamarakshan, V., Das, P., Grosjean, H., Von Delft, F., et al. A small step toward generalizability: training a machine learning scoring function for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 63(10):2960–2974, 2023.

Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S. K., Mer, A. S., Ester, M., and Haibe-Kains, B. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings in Bioinformatics*, 22(6):bbab294, 2021.

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

Svatko, I., Sanchez, M., Bendidi, I., Cottrell, G., and Genovesio, A. Deep learning for bioimaging: What are we learning? *ICLR Workshop on Learning Meaningful Representations of Life*, 2026.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618 (7965):616–624, Jun 2023. doi: https://doi.org/10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9#Sec1.

Toniatti, C., Jones, P., Graham, H., Pagliara, B., and Draetta, G. Oncology drug discovery: planning a turnaround. *Cancer discovery*, 4(4):397–404, 2014.

Tropsha, A., Isayev, O., Varnek, A., Schneider, G., and Cherkasov, A. Integrating qsar modelling and deep learning in drug discovery: the emergence of deep qsar. *Nature Reviews Drug Discovery*, 23(2):141–155, 2024.

Turnbull, O. M., Oglic, D., Croasdale-Wood, R., and Deane, C. M. p-iggen: a paired antibody generative language model. *Bioinformatics*, 40(11):btae659, 2024.

Valsson, Í., Warren, M. T., Deane, C. M., Magarkar, A., Morris, G. M., and Biggin, P. C. Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Communications Chemistry*, 8(1):41, 2025.

Viñas Torné, R., Wiatrak, M., Piran, Z., Fan, S., Jiang, L., Teichmann, S. A., Nitzan, M., and Brbić, M. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, pp. 1–10, 2025.

Volkov, M., Turk, J.-A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y., and Rognan, D. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry*, 65(11):7946–7958, 2022.

Walters, W. P. and Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

Walters, W. P. and Barzilay, R. Critical assessment of ai in drug discovery. *Expert opinion on drug discovery*, 16(9): 937–947, 2021.

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., and Gu, Q. Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*, pp. 52309–52333. PMLR, 2024.

Wang, Z., Zheng, L., Liu, Y., Qu, Y., Li, Y.-Q., Zhao, M., Mu, Y., and Li, W. Onionnet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Frontiers in chemistry*, 9:753002, 2021.

Wenteler, A., Occhetta, M., Branson, N., Curean, V., Huebner, M., Dee, W., Connell, W., Chung, S. P., Hawkins-Hooker, A., Ektefaie, Y., et al. Perteval-scfm: Benchmarking single-cell foundation models for perturbation effect prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Wu, Y., Wershof, E., Schmon, S. M., Nassar, M., Osiński, B., Eksi, R., Yan, Z., Stark, R., Zhang, K., and Graepel, T. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Xiao, Y., Zhao, W., Zhang, J., Jin, Y., Zhang, H., Ren, Z., Sun, R., Wang, H., Wan, G., Lu, P., et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.