Predicting Dose-Response Curves with Deep Neural Networks

Pedro A. Campana [1] **Paul Prasse** [1] **Tobias Scheffer** [1]

Abstract

Dose-response curves characterize the relationship between the concentration of drugs and their inhibitory effect on the growth of specific types of cells. The predominant Hill-equation model of an ideal enzymatic inhibition unduly simplifies the biochemical reality of many drugs; and for these drugs, the widely-used drug performance indicator of the half-inhibitory concentration IC_{50} can lead to poor therapeutic recommendations and poor selections of promising drug candidates. We develop a neural model that uses an embedding of the interaction between drug molecules and the tissue transcriptome to estimate the entire dose-response curve rather than a scalar aggregate. We find that, compared to the prior state of the art, this model excels at interpolating and extrapolating the inhibitory effect of untried concentrations. Unlike prevalent parametric models, it it able to accurately predict dose-response curves of drugs on cells with previously unseen transcriptomes as well as of previously untested drug molecules on established cell lines. Our implementation is available at https://github.com/alonsocampana/ARCANet.

1. Introduction

The concept of a dose-response curve lies at the heart of both drug discovery and personalized medicine: it describes how the growth of a type of tissue is inhibited as a function of the concentration of a drug which the cells are exposed to. Estimates of the dose-response curves of candidate drugs for certain targeted cells—often tumor tissue—as well as for types of healthy tissue are used to predict whether the potential drugs may have a therapeutically useful and safe dosage range. To this end, several large-scale projects perform high-volume dose-response screenings in which the growth-inhibitory properties of different candidate molecules and concentrations are studied *in vitro* for panels of tumoral cells with known genotypes and transcriptomes.

An inhibitory effect $E(x)$ is a rate of decline in the number of living cells after a fixed time interval in which a cell culture has been exposed to a concentration x of a drug compound, where $E(x) = 0$ indicates perfect conservation, and $E(x) = 1$ means total eradication of the cell culture. A dose-response curve describes the effect $E(x)$ as a function of the concentration x. The Hill equation is a widely used parametric dose-response model that assumes a sigmoidal transition from the inhibitory effect E_0 in absence of the drug to the maximal effect of the drug E_∞ at its saturation:

$$\hat{E}_{\text{Hill}}(x) = E_0 + \frac{E_\infty - E_0}{1 + e^{\beta x + \beta_0}}, \tag{1}$$

where E_0, E_∞, β, and β_0 are model parameters. Parameters of multiple drugs and cell lines can be modeled jointly by mixed-effect or hierarchical Bayesian models (Lindstrom & Bates, 1990; Labelle et al., 2019), or they can be estimated independently for each drug-cell pair. In both cases, each drug and cell line has individual parameters that have to be estimated, which renders zero-shot predictions for previously unseen drugs or cell lines impossible.

The Hill equation is an appropriate model of ideal enzymatic inhibition, but it is known to be a simplification of the biochemical reality of most drugs; it is either monotonically increasing or decreasing, and the curve and its derivative are symmetric to its inflection point. The Hill equation cannot adequately describe the behavior of compounds that interact with multiple cellular mechanisms; for instance, it is unable to model the fairly common case of *biphasic* compounds (Mattson, 2008) that have a stimulating effect at a low dose but become toxic at high doses. This model inadequacy cannot be alleviated by increasing amounts of training data. Hill curves are aggregated into a scalar value IC_{50}, the concentration that inhibits half of the cells: $\hat{E}_{\text{Hill}}(IC_{50}) = \frac{1}{2}$. The IC_{50} is frequently used as a performance indicator of drugs. But for drugs whose behavior is poorly approximated by the Hill equation, this value can be misleading, and decisions based on this indicator can result in unnecessary and unsuccessful animal trials, or in failures to detect effective molecules.

[1]Department of Computer Science, University of Potsdam, Potsdam, Germany. Correspondence to: Pedro A. Campana <alonsocampana@uni-potsdam.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

In this paper, we explore the approach to estimate dose-response curves with a universal estimator instead of a parametric model. Since curves are inferred from drug embeddings, cell embeddings, and embeddings of the interaction between genes and the molecule, our model is able to produce zero-shot estimates of curves that are based on the molecular structure of new drugs, the transcriptome of new cell lines, or both. This manuscript makes a number of contributions.

1. We develop the first non-parametric model that estimates dose-response curves in their entirety, as opposed to scalar aggregates such as the half-inhibitory concentration IC_{50} or inhibitory effects at pre-defined concentrations. Its optimization criterion includes a loss that is defined on a variational representation of curves. Being a "biochemistry-informed" universal approximator, the model is able to describe the behavior of drug molecules that deviate from the biochemical model of ideal enzymatic inhibition.
2. We develop the network in such a way that it learns embeddings of gene expressions, drug molecules, and interactions between drugs and gene expressions, which allows the model to produce zero-shot predictions of dose-response curves for unseen drugs and cell lines.
3. For the first time, this paper distinguishes and experimentally studies separate performance measures for the tasks of smoothing, interpolation, and extrapolation of curves as well as precision oncology and drug discovery.
4. We report on experiments that show that the model excels at interpolating and extrapolating curves to new concentrations, at predicting dose-response curves for previously unseen cell lines—which corresponds to the use case of precision oncology—, and at predicting dose-response curves for untried drug molecules—corresponding to the use case of drug discovery. We find that while Morgan fingerprint representations of drugs work best for predictions on known cell lines and drugs, graph representations of drug molecules lend themselves to zero-shot predictions on unseen combinations of cell lines and drugs.

2. Related Work

The Hill equation (Equation 1) is often referred to as *four-parameter (4P) logistic model*. A model that additionally assumes that the inhibitory effect in absence of the drug E_0 is zero is called the *three-parameter (3P) logistic model*. The additional assumption that $E_\infty = 1$ results in the *two-parameter (2P) logistic model* that has a slope and an offset parameter for each pair of drug and cell line (Fallahi-Sichani et al., 2013).

Parameters of multiple drugs and cell lines are usually estimated jointly by mixed-effect (Lindstrom & Bates, 1990) and hierarchical Bayesian (Labelle et al., 2019) models. The well-established *2P mixed-effect logistic model* (Vis et al., 2016) assumes that parameters of a drug and cell line are generated from a global mean parameter vector, plus a mean vector (across drugs) for the cell line that is governed by a multivariate normal distribution with full covariance matrix, plus *iid* noise for each combination. This model still has parameters for each combination, but these parameters are first regularized towards a global population mean, secondly towards the parameters of other cell lines according to the strength of their correlation, and thirdly for each drug towards the cell line mean.

Established models have at least two parameters for each combination of drug and cell line, and therefore cannot make predictions (other than global mean values) for unseen cell lines or unseen drugs. By contrast, Rahman et al. (2019) and Fu et al. (2021) developed adaptations of the random-forest algorithm specifically designed to handle functional data. Both develop node-splitting criteria that use the Kullback-Leibler (KL) divergence between curves and represent cell lines in terms of their gene-expression profiles. Rahman et al. (2019) use various molecule descriptors, Fu et al. (2021) use Morgan descriptors (Capecchi et al., 2020) for each drug as input, enabling the generation of dose-response curves for fixed concentrations. Notably, the model does not explicitly consider the concentrations utilized to produce each response and can only predict the responses for a previously fixed set of concentrations.

Only few other known models extend the model space beyond the logistic Hill equation. Veroli et al. (2015) use products of multiple logistic functions, but their work is not reproducible and their software is only available as an app that processes single dose-response curves in a GUI. Wheeler (2018) model dose-response curves as splines. Their method requires inversions of kernel matrices whose size is quadratic in the number of cell line-drug pairs, which is prohibitive for the data used in our experiments. Tansey et al. (2021) use a neural network to parametrize a Bayesian model as a function of multi-omics cell line features and the identity and concentration of each drug. Since their algorithm identifies and discards outliers in both the training and test data, it only produces predictions for some of the test data and therefore cannot be compared to other methods.

3. Problem Setting

The inhibitory effect $E_{ij}(x)$ of drug i on cell line j at a concentration of x is defined as the expected rate of decline in the number of viable cells after a standard exposure interval, in most cases of 72 hours (Yang et al., 2012). The inhibitory effect is an unknown function, but (generally noisy) obser-

vations $\{(i_k, j_k, x_k, E_{i_k j_k}(x_k))|1 \leq k \leq n)\}$ are available as training data.

We study two different encodings of drug molecules: Morgan fingerprints and annotated graphs. A Morgan fingerprint encodes the atom groups of a chemical compound into a binary vector (Rogers & Hahn, 2010; Capecchi et al., 2020); we employ the rdkit implementation with a radius of 2 and vectors of 2,048 bits. In the graph representation, nodes represent atoms, node annotations reflect the degree of each node, the atom type, the number of neighboring heavy atoms, the formal charge of that atom, the hybridization type, binary variables that indicate whether the atom is contained in a ring and whether it is contained in an aromatic ring, the mass of the atom, its scaled van der Waals radius, and its scaled covalent radius. Edge annotations are the type of edge (single, double, triple or aromatic), and binary variables indicating whether the edge is conjugated or not, and whether the edge is part of a ring. All discrete annotations are one-hot encoded.

Cell lines are represented by the expression levels of genes. In our experiments, we use a set of 2,089 gene that were selected by network propagation in a protein-protein interaction graph (Manica et al., 2019); they correspond to the genes with the highest random-walk probability to the genes targeted by each drug. The objective of the problem of *dose-response curve prediction* is to derive a model $\hat{E}_{ij}(x)$ that approximates $E_{ij}(x)$ as closely as possible for each drug i—including previously unseen drugs—all cell lines j—again, including previously unknown cases—and each concentration x. Model $\hat{E}_{ij}(x)$ therefore has to generalize along several dimensions that we evaluate separately.

1. The standard evaluation of dose-response curves measures the *mean squared error (MSE)* between measured and estimated effect for test combinations of drugs and cell lines that occur in the training data, for concentrations that are missing at random in each curve. It captures the ability of the model to remove noise stemming from the experimental measurement process from each observation, generating a "smooth", less noisy curve in the process. This use case is referred to as *smoothing*.

2. The model's ability to *interpolate to unknown concentrations* within the interval of observed values quantifies its ability to estimate the shape of the curve faithfully. This is evaluated in terms of the MSE for combinations of drugs and cell lines that occur in the training data, but on concentrations that do not occur in the training data for this drug, inside the observed interval of concentrations.

3. Screening experiments can only be performed for a limited number of concentrations, and therapeutically useful doses may turn out to lie outside of the observed range. We measure the model's ability to *extrapolate to higher doses*, outside the interval of observed concentrations. This is evaluated in terms of the MSE for known combinations of drugs and cell lines, but on concentrations that do not occur in the training data for this drug, beyond the observed value range.

4. The model's ability to generalize to unseen cell lines quantifies its merit for the use case of precision oncology, selecting a therapy for a given tumor case at hand. It is measured as MSE on previously unseen cell lines by stratified cross validation along cell lines.

4. Method

Figure 1 shows our *Anti-Cancer Response Curve Analysis Neural Network (ARCANet)* architecture and its functional modules. On a high level, drugs and gene expression levels of cell lines are encoded and combined into a drug-cell line interaction embedding. Together with a drug concentration, this results in a treatment embedding, from which a regression model determines the dose-response curve.

In the *expression encoder*, gene-expression levels of the input cell line are encoded using dropout with probability p_{gene} and a fully-connected layer, followed by a ReLU activation function. For the *drug encoder* and *drug-cell line interaction module*, we consider two alternative designs, shown in Figure 2.

A graph neural network encoder (Figure 2 left) uses m_{mol} residually-connected graph-attention layers with ReLU activation on the molecule graph, followed by m_{full} graph-attention layers on the fully-connected version of the molecule graph to create embeddings of the nodes. Multi-head attention is used to integrate these atom-wise node embeddings and the cell line embeddings into an interaction embedding. With a dropout probability of p_{int}, the attention scores of nodes are set to $-\infty$.

An alternative lower-capacity drug encoder processes the Morgan fingerprint of the drug with a fully-connected layer, followed by a ReLU activation and dropout (Figure 2 right). After both encoders, the use of batch normalization is controlled by a hyper-parameter. A matching version of the drug-cell interaction model concatenates drug and cell line embeddings and employs fully connected layers.

The *treatment model* integrates the drug-cell line embedding and the input concentration into a treatment embedding. The main shortcoming of the Hill equation is that it describes a single enzymatic inhibition, whereas in reality drug molecules may interact with multiple cellular mechanisms, and those interactions may take effect at different concentrations. This motivates the next step in which the

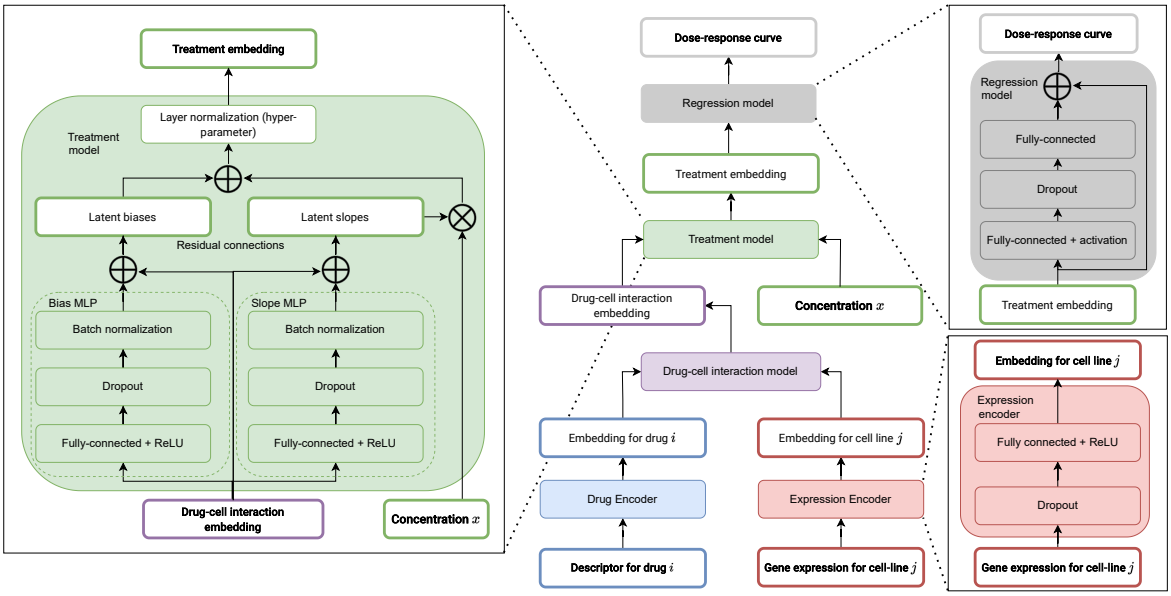

Figure 1. Overview of the architecture of ARCANet and its functional modules.

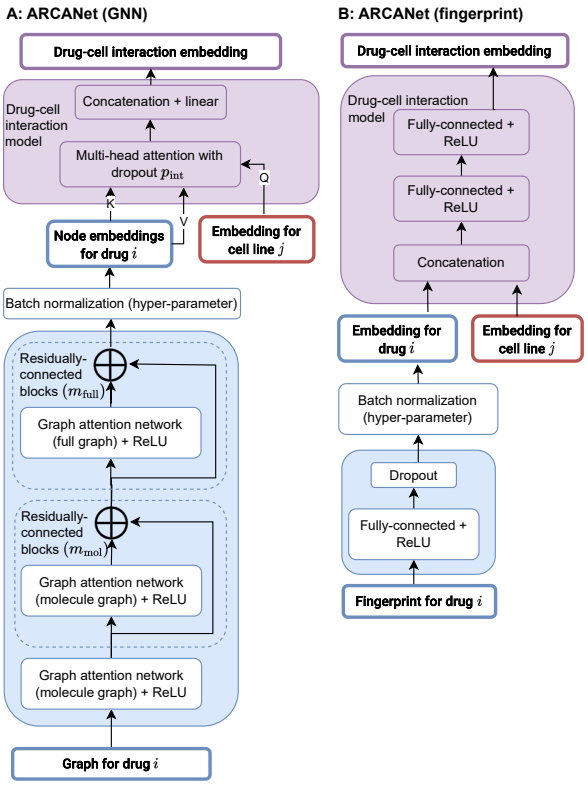

Figure 2. Drug-cell interaction modules for graph (A) and Morgan fingerprint (B) representations of drugs.

interaction embedding is mapped to vectors of slope and bias parameters of Hill curves (resudually connected bias and slope MLP in Figure 1). Element-wise multiplication of concentration and slope parameters, addition of the bias parameter and application of a sigmoid function results in a vector of latent Hill responses—the treatment embedding. The dropout probability p_{Hill} and whether or not to use batch normalization are hyper-parameters of the model. The use of layer normalization after both versions of the treatment models is toggled by a hyper-parameter.

Finally, the *regression model* integrates the treatment embeddings into a dose response curve. The residual connection of the two additional fully-connected hidden layers (the activation function is a hyper-parameter) ensures that the network is biased towards learning a mixture of Hill responses; it is physics-informed because it is regularized towards describing inhibitory effects as a superposition of enzymatic inhibitions. Nevertheless, the fully-connected layers also ensure that it is in fact a universal approximator and can produce any shape of curve, limited only by the tunable number of units.

For the majority of drug-cell line pairs, little or no inhibitory effect can be observed. Some cytotoxic drug molecules, mostly chemotherapies, inhibit almost all cells. When ARCANet is trained with the *mean square error (MSE)* as optimization criterion, parameters that lead to a mean curve per drug, with most curves being mostly flat, are a strong local minimum. Prediction errors for drug molecules that have a targeted inhibitory effect only for some types of tissue at sufficiently high concentrations have a relatively small impact on the overall MSE. However, such targeted molecules have the highest therapeutic merit.

Table 1. Data sets used in the experiments with the number of concentrations, points per curve, drugs, cell lines and data points.

	Concentrations	Points per curve	Drugs	Cell-lines	Data points
CTRPv2	323	16	545	683	4,055,696
NCI60	717	5	47,859	58	12,221,540
GDSC1	111	9	199	941	1,307,070
GDSC2	166	7	153	941	754,152
PRISM	1,042	8	648	479	2,003,384

We approach this issue by defining the loss on a variational representation of observed and estimated dose-response curves. We normalize each curve into a measure that sums to one over all concentrations, and offers control over the penalization of errors on flat curve sections versus errors on curve section with large gradient via a tunable temperature hyper-parameter. Measure $M_E^\tau(x)$ with temperature τ is the result of a softmax operation on a dose-response curve $E(x)$ for which measurements at concentrations $x \in X$ are given:

$$M_E^\tau(x) = \frac{e^{\frac{E(x)}{\tau}}}{\sum_{x' \in X} e^{\frac{E(x')}{\tau}}}. \tag{2}$$

This variational representation allows us to employ the Kullback-Leibler divergence as a loss function on entire dose-response curves:

$$\ell_{KL}(M_E^\tau, M_{\hat{E}}^\tau) = \sum_{x \in X} M_E^\tau(x) \log \frac{M_E^\tau(x)}{M_{\hat{E}}^\tau(x)}, \tag{3}$$

where $E(x)$ is the observed and $\hat{E}(x)$ the predicted dose-response curves, and X is the set of concentrations for which measurements are contained in the training data.

Since the variational representation normalizes all dose-response curves such that it sums to one, information about the absolute magnitude of the inhibitory effects is lost. We reintroduce the absolute effect magnitude by adding the point-wise mean squared error between the observations and the predictions. The total loss $\ell(M_E^\tau, M_{\hat{E}}^\tau)$ amounts to:

$$\begin{aligned} &\ell(M_E^\tau, M_{\hat{E}}^\tau) \\ &\quad = \alpha \ell_{KL}(M_E^\tau, M_{\hat{E}}^\tau) + (1-\alpha)\ell_{MSE}(M_E^\tau, M_{\hat{E}}^\tau). \end{aligned} \tag{4}$$

5. Experiments

This section describes the data, experimental setting, and experimental findings.

5.1. Data

We use the largest publicly available repositories of dose-response data; Table 1 summarizes the data set characteristics. The *Genomics of Drug Sensitivity in Cancer project (GDSC)* (Yang et al., 2012; Iorio et al., 2016) is a precision-oncology data set that contains the largest number of cell lines in any public data set. *Cancer Therapeutics Response Portal (CTRPv2)* (Rees et al., 2015), and *PRISM* (Corsello et al., 2020) are precision-oncology data set with smaller collections of cell lines but larger libraries of drugs. *NCI60* (Shoemaker, 2006) is a dose-response data set for drug discovery; it covers a substantially larger panel of drugs, but fewer cell lines.

We use all dose-response curves for which the drug molecule and the gene-expression profile of the cell line are included in the data; we discard curves that do not have the number of data points described in Table 1.

5.2. Training Environment

All neural network models and baselines are trained, using the PyTorch Geometric library (Fey & Lenssen, 2019) and PyTorch (Paszke et al., 2019) on one NVIDIA A100-SXM4-40GB GPU.

5.3. Evaluation and Hyper-Parameter Tuning

We adapt the evaluation protocol to investigate each of the five dimensions of generalization laid out in Section 3.

1. We evaluate model performance for the use case of *smoothing*. For this, for each curve containing the response of a cell line treated by different dosages of one drug, we split the curve randomly and measure the mean MSE for the data points unseen during training.
2. For *interpolation*, for each drug we split the data over the inner concentration values; the highest and lowest concentrations are always part of the training data. We measure the mean MSE for the drug concentrations unseen during training.
3. For *extrapolation*, we partition the data in a way where each evaluation set contains the highest concentrations of 10% of the drugs, and all the lower concentrations are always part of the training data. We measure the mean MSE for the drug concentrations unseen during training
4. For *precision oncology*, we split the data stratified along cell lines and measure the mean MSE cell lines unseen during training.
5. For *drug discovery*, we split the data stratified along drugs and measure the mean MSE for drugs unseen during training.

We assess the statistical significance of our findings by a two-tailed independent t-test, where we compare both versions of ARCANet against the best-performing baseline.

Table 2. Hyper-parameter search space and identified values.

Hyper-parameter	Search space	Precision oncology on GDSC1	Drug discovery on NCI60
Weight α for loss in Equation 4	[0.00, 1]	0.26	0.57
Learning rate	$[10^{-7}, 10^{-1}]$	3.7×10^{-5}	1.7×10^{-4}
Temperature τ in Equation 2	[0.1, 20]	2.5	2.47
Gradient clipping norm	[0.1, 10]	6.4	2.4
Embedding size	[64, 1024]	624	468
Dropout for drug-cell interaction module (p_{int})	[0, 0.5]	0.31	0.46
Dropout for treatment module (p_{Hill})	[0, 0.5]	0.01	0.26
Dropout for expression encoder (p_{gene})	[0, 0.5]	0.07	0.38
Log-transform concentrations	{True, False}	False	True
Batch normalization in drug encoder	{True, False}	True	True
Batch normalization in treatment model	{True, False}	False	True
Layer normalization in treatment model	{True, False}	False	True
Number of cross-attention heads in drug-cell interaction model	[1, 6]	2	1
Graph attention heads in drug encoder	[1, 4]	4	2
Graph attention layers m_{mol} in drug encoder	[1, 10]	8	6
Graph attention layers m_{full} in drug encoder	[0, 3]	3	1
Fully connected units in regression model	[512, 4096]	2048	1969
Activation function in regression model	{sigmoid, ELU, ReLU, tanh}	ReLU	tanh

We optimize hyper-parameters using the Bayesian optimization with 100 configuration proposals and early stopping using the median stopping rule in the Optuna (Akiba et al., 2019) framework. PRISM, GDSC2 and CTRPv2 have been collected with a precision-oncology application focus in mind whereas NCI60 is a drug-discovery data set. NCI60 measures inhibitory effects using a different experimental methodology which leads to different numerical values and potentially different optimal hyper-parameter values. NCI60 also contains more drugs by two orders of magnitude. We therefore run one hyper-parameter tuning processes jointly for PRISM, GDSC2 and CTRPV2 and a separate process for NCI60. Both the search space and the final configuration are shown in Table 2.

5.3.1. PRISM, GDSC2 AND CTRPv2

GDSC consists of the older GDSC1 and the newer GDSC2 subsets. The two subsets report observations that were made in independent experiments. We reserve the older GDSC1 data exclusively for prototyping and hyper-parameter tuning, using 90% of the cell lines for training and 10% for hyper-parameter tuning and the MSE for precision oncology as objective. We use GDSC2, CTRPv2, and PRISM, respectively, for training and evaluation.

We split each of these data sets separately for smoothing, interpolation, extrapolation, and precision oncology, as described in Section 5.3. We partition the available data into 10 folds, unless a data set has fewer points per curve for smoothing and interpolation, in which case we adjust the number of folds. We finally perform cross-validation over these folds, training each model for 100 epochs.

5.3.2. NCI60

For NCI60, we use a single three-way split into 80% of the drugs for training, 10% of drugs for hyper-parameter tuning, and 10% for evaluation; we limit the number of training epochs to 50 and use the MSE for drug discovery as optimization criterion. After hyper-parameter optimization, we average 10 iterations of training and evaluation using random inizializations and permutations of the training and tuning data, but holding out the same 10% of unseen evaluation data.

5.4. Reference Methods

We compare ARCANet to the state of the art in dose-response modeling.

1. We apply individual 2P, 3P and 4P models (see Section 2) fitted to each dose-response curve individually. This model cannot be applied for previously unseen cell lines or drugs.

2. We apply the *2P mixed-effect logistic model* (Vis et al., 2016) in the implementation used by the GDSC project (Yang et al., 2012; Vis et al., 2016; git). This model cannot be applied for previously unseen cell lines or drugs.

3. We use the functional random forest *(FunFor)* of Fu et al. (2021) because it is more recent, and it considers the competing model of Rahman et al. (2019) as a baseline. This model makes predictions for a fixed set of concentrations and cannot be applied for smoothing, interpolation, or extrapolation.

5.5. ARCANet Outperforms the State of the Art for Smoothing, Interpolation, and Extrapolation

In this section, we study the performance of ARCANet and reference models for smoothing, interpolation, and extrapolation. Table 3 shows that for smoothing, both ARCANet (GNN) and ARCANet (fingerprint) outperform all reference

Table 3. MSE (standard errors in parentheses) for smoothing, interpolation, and extrapolation. "×" indicates non-convergence. The best result is highlighted in bold, statistically significant improvements of ARCANet over the best-performing reference method ($\alpha \leq 0.05$) are marked "*".

	Smoothing				Interpolation				Extrapolation			
	CTRPv2	GDSC2	NCI60	PRISM	CTRPv2	GDSC2	NCI60	PRISM	CTRPv2	GDSC2	NCI60	PRISM
2P (individual)	0.069 (0.0)	0.012 (0.0)	0.099 (0.0)	2.165 (0.004)	0.055 (0.002)	0.009 (0.001)	0.051 (0.0)	2.308 (0.11)	0.303 (0.013)	0.026 (0.005)	0.309 (0.001)	3.697 (0.214)
3P (individual)	0.022 (0.0)	0.005 (0.0)	0.159 (0.0)	1.055 (0.003)	0.013 (0.0)	0.004 (0.001)	0.049 (0.0)	1.166 (0.049)	0.127 (0.007)	0.015 (0.004)	0.591 (0.002)	1.957 (0.165)
4P (individual)	0.023 (0.0)	0.006 (0.0)	0.141 (0.0)	1.154 (0.003)	0.013 (0.0)	0.004 (0.001)	**0.045** (0.0)	1.234 (0.063)	0.116 (0.006)	0.012 (0.002)	0.528 (0.002)	1.503 (0.107)
2P (Vis et al., 2016) (mixed effect)	×	0.005 (0.0)	×	×	×	0.004 (0.0)	×	×	×	0.011 (0.002)	×	×
ARCANet (GNN)	0.019* (0.0)	**0.003*** (0.0)	0.057* (0.008)	0.335* (0.005)	0.019 (0.0)	0.004 (0.0)	0.086 (0.012)	0.784* (0.038)	0.064* (0.003)	0.011 (0.002)	0.247* (0.038)	1.067* (0.058)
ARCANet (fingerprint)	**0.011*** (0.0)	**0.003*** (0.0)	**0.047*** (0.003)	**0.271*** (0.001)	**0.012*** (0.0)	**0.003** (0.0)	0.051 (0.004)	**0.68*** (0.026)	**0.053*** (0.002)	**0.008** (0.001)	**0.201*** (0.004)	**0.854*** (0.04)

methods significantly ($p \leq 0.05$) for all data sets. ARCANet (fingerprint) always performs better than or as good as ARCANet (GNN). For interpolation, ARCANet (fingerprint) has the lowest MSE for three of the four data sets with significant improvements in two cases; only for NCI60, the 4P baseline has the lowest MSE, but the improvement over ARCANet (fingerprint) is not significant. For extrapolation, ARCANet (fingerprint) exhibits the lowest MSE; the improvement over all reference methods is significant for three out of four data sets for both versions of ARCANet. The 2P mixed effects model (Vis et al., 2016) makes use of restricted maximum likelihood and can become numerically unstable when the model is misspecified for the data. It only converges for GDSC2.

For smoothing, interpolation, and extrapolation, predictions are made for combinations of cell lines and drugs that are represented in the training data, which renders the model's ability do generalize across drug compounds less crucial and gives an advantage to the lower-capacity version of ARCANet that represents drugs by Morgan fingerprints.

5.6. ARCANet Outperforms the State of the Art for Precision Oncology and Drug Design

This section explores the performance of ARCANet and reference models for zero-shot predictions on unseen cell lines and unseen drugs. Table 4 shows that both versions of ARCANet significantly outperform functional random forests (Fu et al., 2021) on unseen cell lines for precision oncology, and ARCANET (GNN) significantly outperforms functional random forests on unseen drug compounds for drug discovery. Note that none of the parametric reference methods can be applied in this scenario because they have parameters for each cell line and drugs.

In these settings, the model's ability to learn an internal representation of drugs that is conducive to generalization across unseen combinations of cell lines and drugs is crucial. In this setting, the graph representation and graph-attentional layers show an advantage over the Morgan fingerprint representation of compounds.

Table 4. MSE (standard errors in parentheses) for precision oncology and drug discovery. The best result is highlighted in bold, statistically significant improvements of ARCANet over the best-performing reference method ($\alpha \leq 0.05$) are marked "*".

	Precision oncology			Drug discovery
	CTRPv2	GDSC2	PRISM	NCI60
FunFor (Fu et al., 2021)	0.063 (0.001)	0.009 (0.0)	1.185 (0.019)	0.14 (0.001)
ARCANet (GNN)	**0.03*** (0.0)	**0.005*** (0.0)	0.425* (0.006)	**0.12*** (0.005)
ARCANet (fingerprint)	0.031* (0.001)	**0.005*** (0.0)	**0.422*** (0.005)	0.13 (0.005)

5.7. Parametric Models Are Best for Very Small Samples

ARCANet is a high-capacity neural network whereas the parametric reference models only have between 2 and 4 parameters for each combination of cell line and drug. This section therefore studies how the relative performance of these models depends on the available training sample size in the smoothing setting on the GDSC2 data. Figure 3 shows that the performance of ARCANet only falls short of the performance of the 3P and 4P models when the number of training data points is reduced by roughly two orders of magnitude. The low-capacity parametric reference models are highly robust against training-data scarcity.

5.8. Ablation Study

This section studies the impact of removing three elements of the ARCANet architecture. The first ablation replaces

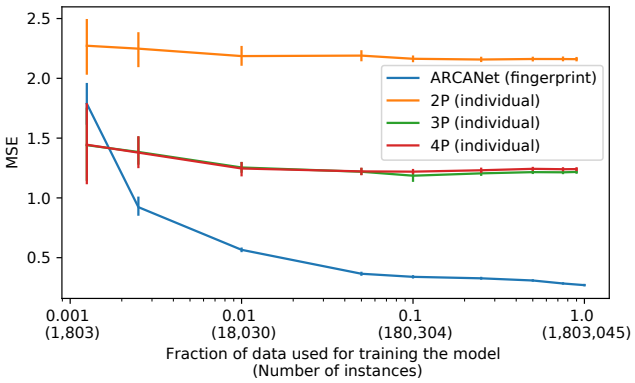

Figure 3. MSE over the fraction and absolute number of training data points; error bars show the 95% confidence interval. Training data are shrunk by deleting measurements for random combinations of drugs, cell lines, and concentrations.

Table 5. MSE (standard errors in parentheses) of ablated versions of ARCANet for precision oncology and smoothing.

	Treatment model	$\alpha = 0$ (Eq. 4)	Linear regr. model	Precision oncology	Smoothing
ARCANet (fingerprint)	Hill	×	×	0.422 (0.005)	0.271 (0.001)
	Concat.	×	×	0.423 (0.006)	0.272 (0.0)
	Hill	×	✓	0.431 (0.006)	0.281 (0.001)
	Hill	✓	×	0.427 (0.005)	0.273 (0.001)
ARCANet (GNN)	Hill	×	×	0.425 (0.006)	0.335 (0.005)
	Concat.	×	×	0.434 (0.008)	0.337 (0.004)
	Hill	×	✓	0.422 (0.006)	0.327 (0.004)
	Hill	✓	×	0.435 (0.009)	0.329 (0.003)

the mixture of Hill equations in the treatment model by a baseline that concatenates the drug-cell line embedding and the concentration, and applies an MLP equivalent to the bias and slope MLPs to generate the treatment embedding. The second ablation sets α in Equation 4 to zero, thus removing the variational component of the loss function. The final ablation replaces the regression module by a fully connected layer whose number of units is a tunable hyper-parameter. We study these ablations for the smoothing and precision oncology settings on the PRISM data.

Table 5 shows that for ARCANet (fingerprint), all ablations lead to a higher MSE. For the combination of ARCANet (GNN) and the smoothing setting, the ablations do not perform uniformly worse, which is consistent with our earlier finding that the capacity of ARCANet (GNN) is higher than necessary when no generalization across drugs is required. For precision oncology, the ablations of ARCANet (GNN) perform worse, with one statistically insignificant exception for the ablated regression model.

6. Discussion and Conclusion

We have developed ARCANet, the first deep neural-network model that estimates dose-response curves. By estimating dose-response curves in their entirety and without assuming a parametric model, ARCANet is able to describe the behavior of drug molecules that deviate from the biochemical model of ideal enzymatic inhibition. Figure 4 in Appendix A shows two illustrative examples of protein-kinase inhibitors (Rashkov et al., 2016) that display biphasic dose-response curves by stimulating cellular growth at lower concentrations and inhibiting it at higher dosages.

By learning embeddings of gene expressions, graph-attentional embeddings of drug molecules, and cross-attentional embeddings of interactions between drug molecules and gene expressions, ARCANet is able to produce zero-shot predictions of dose-response curves for unseen drugs and cell lines.

Previous work evaluates dose-response predictions in terms of the MSE for cell lines, drugs, and concentrations that also occur in the training data. We argue that in order to evaluate models as a tool for precision oncology and drug discovery, the interpolation of intermediate concentrations, extrapolation to higher than previously observed concentrations, dose responses for previously unseen cell lines, and of previously untried drug compounds need to be evaluated. We can conclude that ARCANet with Morgan fingerprint representations of drug compounds excels at interpolating and extrapolating curves to untried and potentially unsafe drug dosages (Subramanian et al., 2017). While ARCANet with graph representations also outperforms the parametric reference models, it falls short of the lower-capacity version that uses molecular fingerprints when no generalization to unseen combinations of cell lines and drugs is required. For predicting dose-response curves for previously unseen cell lines (precision oncology) and unseen drug compounds (drug discovery), ARCANet with graph representation improves substantially on functional random forests.

The prevailing parametric Hill models cannot be applied to unseen cell lines or drugs. Parametric models beat ARCANet for predictions on previously-seen cell lines and drugs only after the training sample size is reduced by roughly two orders of magnitude, down to a few thousand measurements.

7. Data and Code Availability

All experiments performed are based on publicly available datasets. The code used for downloading and preprocessing the data as well as for reproducing our experiments can be found in `https://github.com/alonsocampana/ARCANet`.

Impact Statement

This paper presents work whose goal, apart from advancing the field of Machine Learning, is to advance the treatment of cancer and potentially other diseases. While the advancement of personalized medicine has an enormous potential for societal benefit, one should be aware of the potential for unethical use. Any method that is effective at identifying inhibitors for cells with a specific transcriptome could potentially also be used to identify highly toxic substances (Urbina et al., 2022), or even substances that are selectively toxic for a targeted population that shares certain genetic commonalities. Such misuse might be addressed by subjecting research on collections of genetic and toxicological data to ethical reviewing.

References

GitHub - CancerRxGene/gdscIC50: An R package to fit dose response curves for data from the Genomics of Drug Sensitivity of Cancer (GDSC) project. — github.com. https://github.com/CancerRxGene/gdscIC50. [Accessed 17-May-2023].

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework, 2019. URL https://arxiv.org/abs/1907.10902.

Capecchi, A., Probst, D., and Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):1–15, 2020.

Corsello, S. M., , et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, January 2020. doi: 10.1038/s43018-019-0018-6. URL https://doi.org/10.1038/s43018-019-0018-6.

Fallahi-Sichani, M., Honarnejad, S., Heiser, L. M., Gray, J. W., and Sorger, P. K. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nature Chemical Biology*, 9(11):708–714, 2013.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fu, G., Dai, X., and Liang, Y. Functional random forests for curve response. *Scientific Reports*, 11(1):24159, 2021.

Iorio, F., Knijnenburg, T., Vis, D., et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016. doi: doi:10.1016/j.cell.2016.06.017.

Labelle, C., Marinier, A., and Lemieux, S. Enhancing the drug discovery process: Bayesian inference for the analysis and comparison of dose–response experiments. *Bioinformatics*, 35(14):i464–i473, 2019.

Lindstrom, M. J. and Bates, D. M. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pp. 673–687, 1990.

Manica, M., Oskooei, A., Born, J., Subramanian, V., Sáez-Rodríguez, J., and Martínez, M. R. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 16(12):4797–4806, 2019. URL https://doi.org/10.1021/acs.molpharmaceut.9b00520.

Mattson, M. P. Hormesis defined. *Ageing Research Reviews*, 7(1):1–7, 2008. URL https://doi.org/10.1016/j.arr.2007.08.007.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Rahman, R., Dhruba, S. R., Ghosh, S., and Pal, R. Functional random forest with applications in dose-response predictions. *Scientific Reports*, 9(1), February 2019. doi: 10.1038/s41598-018-38231-w. URL https://doi.org/10.1038/s41598-018-38231-w.

Rashkov, P., Barrett, I. P., Beardmore, R. E., Bendtsen, C., and Gudelj, I. Kinase inhibition leads to hormesis in a dual phosphorylation-dephosphorylation cycle. *PLOS Computational Biology*, 12(11):e1005216, November 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005216. URL http://dx.doi.org/10.1371/journal.pcbi.1005216.

Rees, M. G., Seashore-Ludlow, B., Cheah, J. H., Adams, D. J., Price, E. V., Gill, S., Javaid, S., Coletti, M. E., Jones, V. L., Bodycombe, N. E., Soule, C. K., Alexander, B., Li, A., Montgomery, P., Kotz, J. D., Hon, C. S.-Y., Munoz, B., Liefeld, T., Dančík, V., Haber, D. A., Clish, C. B., Bittker, J. A., Palmer, M., Wagner, B. K., Clemons, P. A., Shamji, A. F., and Schreiber, S. L. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature Chemical Biology*, 12(2):109–116, December 2015. ISSN 1552-4469. URL http://dx.doi.org/10.1038/nchembio.1986.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10): 813–823, October 2006. doi: 10.1038/nrc1951. URL https://doi.org/10.1038/nrc1951.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O. M., Piccioni, F., Johnson, S. A., Lyons, N. J., Berger, A. H., Shamji, A. F., Brooks, A. N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D. Y., Hu, R., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Greenside, P., Gray, N. S., Clemons, P. A., Silver, S., Wu, X., Zhao, W.-N., Read-Button, W., Wu, X., Haggarty, S. J., Ronco, L. V., Boehm, J. S., Schreiber, S. L., Doench, J. G., Bittker, J. A., Root, D. E., Wong, B., and Golub, T. R. A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017. doi: 10.1016/j.cell.2017.10.049. URL https://doi.org/10.1016/j.cell.2017.10.049.

Tansey, W., Li, K., Zhang, H., Linderman, S. W., Rabadan, R., Blei, D. M., and Wiggins, C. H. Dose–response modeling in high-throughput cancer drug screenings: an end-to-end approach. *Biostatistics*, 23(2):643–665, January 2021. doi: 10.1093/biostatistics/kxaa047. URL https://doi.org/10.1093/biostatistics/kxaa047.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

Veroli, G. Y. D., Fornari, C., Goldlust, I., Mills, G., Koh, S. B., Bramhall, J. L., Richards, F. M., and Jodrell, D. I. An automated fitting procedure and software for dose-response curves with multiphasic features. *Scientific Reports*, 5(1), October 2015. doi: 10.1038/srep14701. URL https://doi.org/10.1038/srep14701.

Vis, D. J., Bombardelli, L., Lightfoot, H., Iorio, F., Garnett, M. J., and Wessels, L. F. Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics*, 17(7):691–700, 2016.

Wheeler, M. W. Bayesian additive adaptive basis tensor product models for modeling high dimensional surfaces: an application to high-throughput toxicity testing. *Biometrics*, 75(1):193–201, August 2018. doi: 10.1111/biom.12942. URL https://doi.org/10.1111/biom.12942.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., and Garnett, M. J. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, November 2012. doi: 10.1093/nar/gks1111. URL https://doi.org/10.1093/nar/gks1111.

A. Example Curves

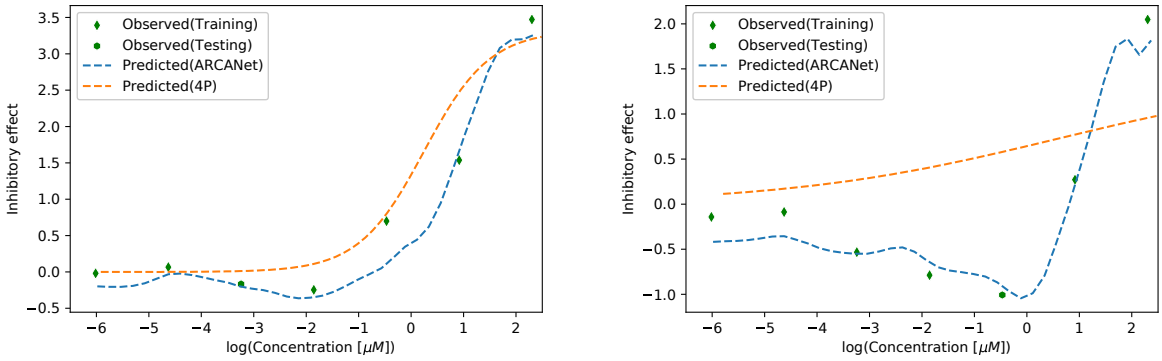

Figure 4. Example dose-responses that deviate from the Hill model. (a) Cell-line ACH-000026 treated with Tyrphostin-A9 (tyrosine kinase inhibitor). (b) Cell-line ACH-000822 treated with AZD2858 (glycogen synthase kinase-3 inhibitor).