Interpretable deep learning architectures for improving drug response prediction: myth or reality?

Yihui Li[1], David Earl Hostallero[1,2], and Amin Emad[1,2,3,*]

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada

[2] Mila, Quebec AI Institute, Montreal, QC, Canada

[3] The Rosalind and Morris Goodman Cancer Institute, Montreal, QC, Canada

* Corresponding Author:

Amin Emad

755 McConnell Engineering Building

3480 University Street, Montreal, Quebec, Canada, H3A 0E9

Email: amin.emad@mcgill.ca

Abstract

Recent advances in deep learning model development have enabled more accurate prediction of drug response in cancer. However, the black-box nature of these models still remains a hurdle in their adoption for precision cancer medicine. Recent efforts have focused on making these models interpretable by incorporating signaling pathway information in model architecture. While these models improve interpretability, it is unclear whether this higher interpretability comes at the cost of less accurate predictions, or a prediction improvement can also be obtained. In this study, we comprehensively and systematically assessed four state-of-the-art interpretable models developed for drug response prediction to answer this question using three pathway collections. Our results showed that models that explicitly incorporate pathway information in the form of a latent layer perform worse compared to models that incorporate this information implicitly. Moreover, in most evaluation setups the best performance is achieved using a simple black-box model. In addition, replacing the signaling pathways with randomly generated pathways shows a comparable performance for the majority of these interpretable models. Our results suggest that new interpretable models are necessary to improve the drug response prediction performance. In addition, the current study provides different baselines models and evaluation setups necessary for such new models to demonstrate their superior prediction performance.

Introduction

Deep learning (DL) models have recently found important applications in the biomedical field including medical imaging, disease diagnosis, patient survival prediction [1, 2], and drug response prediction (DRP) [3-7]. In spite of their success in their perspective tasks, most DL models are considered as "black-boxes" with inner operations that are difficult to interpret. This characteristic of DL models is undesirable for applications in the biomedical field, as identifying the set of biological features that contribute to the model prediction outputs and understanding the relationship between these features are crucial when conducting further experimental studies to validate these computational findings. To address these challenges, the concept of interpretable artificial intelligence [8-10] has been introduced to create models that can achieve both high performance and interpretability.

In the context of DRP, model interpretability can be achieved in two ways: 1) using post-hoc analysis to determine feature attributions and identify important features, and 2) integrating prior knowledge (e.g., signaling pathways) to add meaningful structure to the model, which can then be interpreted. While others and us have successfully used the former strategy in DRP [4, 11, 12] and other applications [13, 14], the latter strategy allows the interpretability to go one step further to provide systems biology insights regarding the mechanisms involved in response to drug treatments. The models that incorporate pathway information have resulted valuable insights regarding drugs' mechanisms of action and gene-pathway relationships, some of which have been validated experimentally [15]. However, there have been conflicting reports on their ability in providing accurate drug response predictions. Ideally, interpretability should not come

at the expense of prediction performance, since a lower prediction performance may reflect that some of the biological nuances captured by the black-box model cannot be recaptured using the more interpretable model.

Recognizing the intertwined relationship between interpretability and performance, the majority of recent models that incorporate pathway information for better interpretability have also sought and reported an improved prediction performance [16-20]. On the other hand, some studies have reported comparable or slightly worse model performance after incorporating pathway information [15]. However, it is rather difficult to gauge the (potential) contribution of pathway information in DRP performance from the original studies, due to differences between data used in each study, their evaluation setup, and in many cases a lack of appropriate baseline models to act as control. To investigate these inconsistent findings in state-of-the-art models, we conducted a study that comprehensively evaluates the effect of pathway incorporation on performance of DRP models and aims to answer five main questions:

1. Does the inclusion of biological pathway information improve model performance when evaluated strictly and comprehensively?
2. Which type of pathway incorporation strategy is best capable of improving the performance?
3. Are interpretable models better suited for prediction of response of unseen cell lines or unseen drugs?
4. Can the performance of the interpretable models be attributed to biological information present in the pathway datasets, or a similar improvement can be also achieved through

the use of randomly generated pathways, reflecting a technical (instead of a biological) origin for the performance?

5. What pathway database is most helpful in improving model performance?

To answer the proposed questions, we performed 171 experiments evaluating 19 computational models with three pathway collections [21], [22], [23] and under three data splitting strategies. The models included four state-of-the-art interpretable DL architectures that incorporate pathway information [16, 17, 19, 20] (henceforth pathway-based models) and four variants of them, as well as eleven baseline models that can capture different types of biases in the analysis. We selected these interpretable models since they use similar type of information for CCLs and drugs and utilize gene-pathway membership in their architectures, allowing us to compare them fairly and comprehensively. Moreover, they represent two categories of strategies to incorporate pathway information in DL architectures: methods that use a pathway layer connecting genes to pathway nodes (*explicit* models) such as PathDNN [16] and ConsDeepSignaling (CDS) [20], and those that do not directly use a pathway layer (*implicit* models) such as HiDRA [17], and PathDSP [19].

Our baseline models included a black-box model with a similar architecture to those of the interpretable models, as well as “naïve” predictors and “random-pathway” predictors, two important baselines that have been largely overlooked in previous studies. The naïve predictor uses the average drug response of samples in the training set and reports that for each testing sample. This baseline is particularly important in controlling for inflation of prediction

performance due to distinct range of log IC50 (natural log of the half maximal inhibitory concentration, a drug sensitivity measure) of different drugs. In other words, it is possible to obtain a good approximation of drug response by simply knowing the identity of the drug, resulting in artificially inflated performance metrics. Each random-pathway predictor exactly matches the architecture and pipeline of an interpretable model, but randomly assigns genes to pathways, while preserving the size of each pathway. These baselines allow us to determine whether potential performance improvement of an interpretable model is truly due to the added value of the biological information, or instead is a technical artifact of modifying the model architecture.

Our analysis showed that overall, incorporating pathway information does not lead to improved prediction performance, confirming the observations reported by Kuenzi et al. [15] for their proposed model. In particular, in many cases a simple black-box multilayer perceptron (MLP) achieves the best performance. Moreover, even in instances that performance improvement compared to an MLP or a naïve predictor was observed, a similar performance was achieved using randomly generated pathways. This suggests that such improvements should not be attributed to the biological information carried by pathway collections and is likely a technical artifact. We also observed that the strategy used to include pathway information in the models has a significant influence on the performance, and explicit models seem to perform worse compared to implicit models. Finally, Reactome pathways seemed to provide slightly better predictions compared to other pathway collections.

108 **Methods**

109 **Data preprocessing and uniform dataset formation**

110 To form uniform datasets for our analyses, we first evaluated different data modalities and
111 datasets used by each of the pathway-based models (Supplementary Table S1). In these studies,
112 gene expression (GEx), somatic mutation (Mut), and copy number variation (CNV) of samples
113 were used, while for drugs their targets (T) or their Morgan fingerprints (FP) capturing their
114 chemical structure were used. In order to maintain fairness and consistency of model
115 performance comparisons, for each choice of pathway collection we compiled a uniform dataset
116 that was used by all models evaluated in this study (three uniform datasets in total). These
117 datasets are freely available in https://zenodo.org/record/7101665#.YzS79HbMKUk.

118

119 We collected GEx, Mut, CNV, and drug sensitivity data (in the form of log IC50) of 959 cancer cell
120 lines (CCLs) from Genomics of Drug Sensitivity in Cancer (GDSC) [24] database. We obtained drug
121 target information from STITCH [25] and drug structural data from PubChem [26]. Protein-protein
122 interactions (PPI) that were used by one of the models were obtained from the STRING database
123 [27] (only experimental PPIs were used). Finally, gene-pathway membership information was
124 obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) [21], PID (Pathway Interaction
125 Database) [22], and Reactome [23]. Supplementary Table S2 outlines the data used in this study
126 and their sources. We obtained drug response data in the form of log IC50 values and removed
127 duplicate drugs that came from different experimental batches. In such cases, we kept the drug
128 whose response was measured across a larger number of CCLs. We collected drug InChI
129 (International Chemical Identifier) strings [28] from PubChem and used the RDKit [29] software

130 to generate 512-bit Morgan fingerprints for these drugs. We obtained drug target data from the
131 STITCH database, where we only kept drug targets with confidence score larger than 800 (out of
132 1000) and coming from the "experimental" and "database" channels.
133
134 We performed log2(FPKM+1) normalization on the GEx data and removed genes whose
135 expression showed low variability across different CCLs (standard deviation < 0.1). We also
136 removed genes for which there were no somatic mutations, CNV, pathway information, drug
137 target data, and STRING Experimental PPI information. This formed our common gene set (Figure
138 1A and 1B). In parallel, drug targets that were not present in the common gene set above or in
139 the PPI network were excluded. Only drugs that had both log IC50 measurements and drug
140 targets were kept in the final uniform datasets (Figure 1C and 1D). The PPI network was involved
141 in the data preprocessing step as PathDSP [19] incorporated it to perform pathway enrichment
142 analysis. Since we needed the uniform datasets to be usable by all models, we included this step
143 in the pre-processing procedure. Figure 1 illustrates the process of constructing the pathway-
144 specific uniform datasets. Since each source of pathway collection contained different number
145 of genes, the final dataset for each collection was slightly different. Table 1 summarizes the
146 number of CCLs, drugs, genes, and pathways for each pathway collection in the uniform dataset,
147 while Supplementary Table S3 provides details about included CCLs and drugs.
148
149 **Table 1:** Summary of pathway-specific uniform datasets.

Pathway Database	Num. CCLs	Num. Drugs	Num. Unique Drug Targets	Num. (Drug, CCL) pairs	Num. Pathways	Num. Unique Genes
KEGG	959	162	446	118,896	332	5511
PID	959	153	321	112,781	196	2078
Reactome	959	177	493	128,324	1608	7831

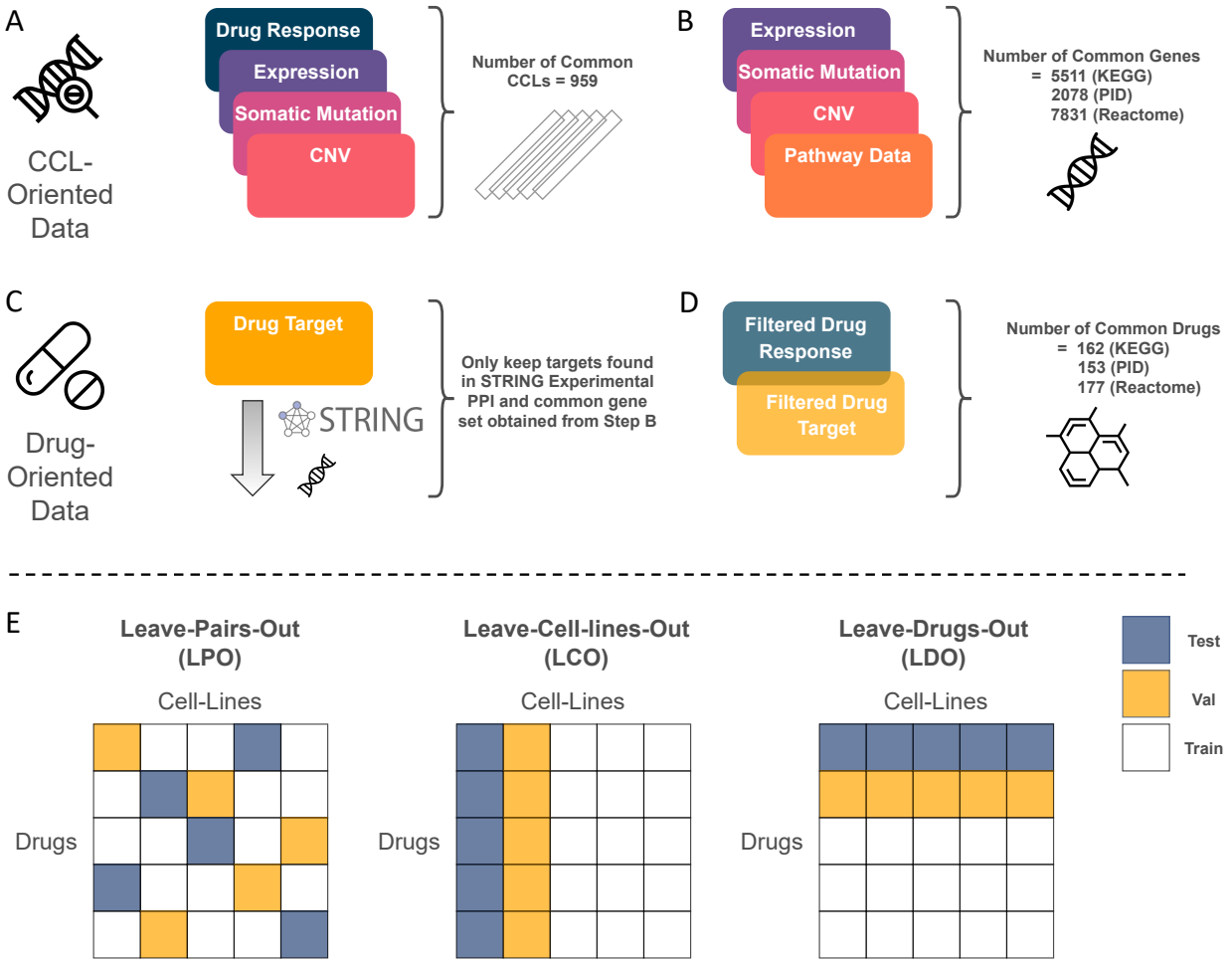

Figure 1: Construction of pathway-specific uniform datasets and data splitting approaches. (A) CCLs with available data for drug response, gene expression, somatic mutation, and CNV were selected. (B) Genes shared between different sources of data were identified. Genes that were not present in any pathway were removed. (C) Drug target genes that were not found in the common gene set obtained from Step B and the STRING Experimental PPI network were removed. (D) Drugs and small molecules that had measured log IC50 values and drug target information were selected. E) Model input data was split into five folds, with the training, validation, and test set ratio of 3: 1: 1. Folds in the LPO validation scheme are formed by randomly selecting mutually exclusive (CCL, drug) pairs, whereas in LCO and LDO validation schemes, mutually exclusive cell lines and drugs are randomly selected, respectively.

163 **Model evaluation and data split**

164 We split our data randomly into five disjoint folds, where the training, validation, and test ratio
165 was 3: 1: 1. The validation set was used for hyperparameter tuning and the test set was used for
166 final model evaluation. The details of hyperparameter tuning, model training, and final
167 architectures are provided in Supplementary File S2. We adopted three data splitting methods
168 (validation schemes) to generate these folds: leave-pairs-out (LPO), leave-cell-lines-out (LCO),
169 and leave-drugs-out (LDO), as depicted in Figure 1E. These three strategists were adopted to
170 comprehensively assess the models for different drug response prediction tasks (for unseen (CCL,
171 drug) pairs, unseen CCLs, and unseen drugs, respectively), and to determine in which one of these
172 tasks (if any) pathway incorporation improves model prediction performance. To ensure fairness,
173 same folds were used for all models.

174

175 We evaluated the performance of each model using two main performance measures:
176 Spearman's correlation coefficient (SCC) and root mean squared error (RMSE). First, for a fixed
177 CCL, the predicted values across all drugs of the test set were compared with the measured log
178 IC50 values to calculate a CCL-specific performance measure (SCC or RMSE). Then, the mean and
179 standard deviation of the performance measure was calculated across all CCLs.

180

181 **Overview of interpretable models and their variants**

182 To study the effect of incorporating pathway information on drug response prediction, we
183 selected four pathway-based state-of-the-art models: PathDNN [16], ConsDeepSignaling (CDS)
184 [20], HiDRA [17], and PathDSP [19]. We selected these models since 1) they use similar types of

185 information for CCLs and drugs and utilize gene-pathway membership in their architectures
186 (instead of other types of prior information such as hierarchical relationships of gene ontologies),
187 2) they all showed improved drug response prediction performance in their original studies
188 compared to their black-box counterparts, and 3) they represent two important categories of
189 implicit and explicit models (as discussed earlier). While other important models also exist (e.g.,
190 DrugCell [15]), they did not satisfy the conditions above. For example, DrugCell (unlike the models
191 above) uses the hierarchical structure of gene ontologies and pathways, making it rather difficult
192 to compare against the models above in a fair manner, since it takes advantage of more detailed
193 information. Moreover, the original study of DrugCell showed that while including prior
194 information improved interpretability of their model, it did not improve the performance of drug
195 response prediction compared to its black-box counterpart. Due to the reasons above, we
196 decided to exclude it from this analysis.
197
198 Models with an *explicit* pathway layer (e.g., PathDNN and CDS) typically define a gene and a
199 pathway layer with connections between these layers reflecting gene-pathway membership
200 (Figure 2). The input layer of this category of models contains drug and cell line features at gene
201 level. As a result, only drug gene targets can be used with these models and Morgan Fingerprint
202 (and other structural data) is not usable without altering the model architecture. The pathway
203 layer is then connected to a group of fully-connected layers to predict drug response for a given
204 sample. The inclusion of the pathway layer allows identification of important pathways for a
205 particular drug treatment or cancer type through post-hoc feature importance analysis.
206

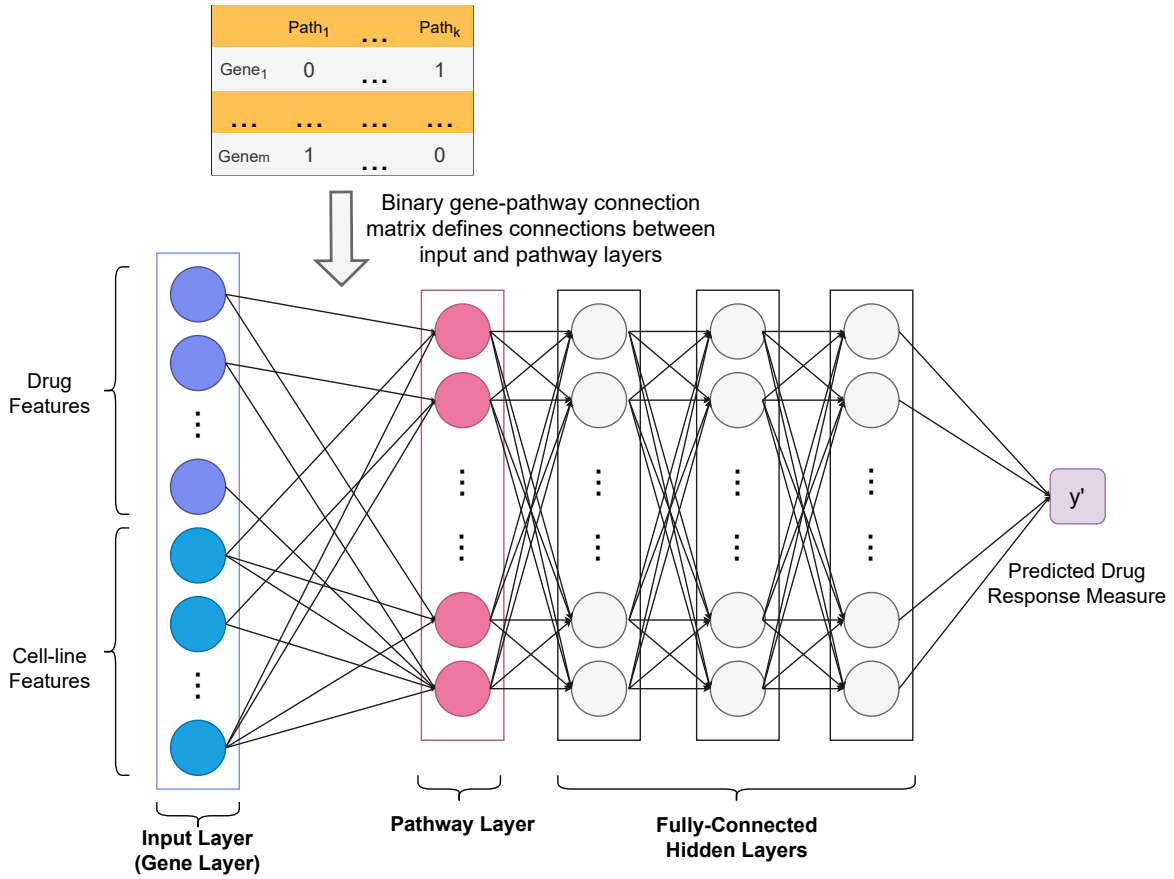

Figure 2: An overview of explicit pathway models using pathway and gene layers. In the first hidden layer (pathway layer), each node represents a pathway. A binary gene-pathway connection matrix (0 = no connection, 1 = connection) defines connections between the gene and pathway layer based on pathway membership. The pathway layer is then connected to a group of fully-connected layers to produce the final drug response prediction.

Models that *implicitly* incorporate pathway information take various forms. For example, HiDRA [17] uses a gene-level and pathway-level attention module to calculate pathway importance scores, where a small-scale neural network is dedicated to each pathway by only using features associated with the member genes of that specific pathway as inputs. On the other hand, PathDSP uses a classic fully connected feedforward architecture, but the input features are pathway-enrichment scores rather than gene-level features. See Supplementary Files S1 and S2 for details regarding models' architectures and their training procedure.

Table 2: List of evaluated models. △ = universal baseline, ◊ = model-specific random pathway baseline, ★ = original pathway-based model, ■ = model variant, GEx = gene expression, CNV = copy number variation, Mut = somatic mutation, T = drug target data, FP = Morgan fingerprint (drug structural data)

Model Name	Model Variant Name	Cell Line Features	Drug Features
Five-Layer MLP	MLP (GEx, FP) △	GEx	FP
	MLP (GEx, T) △	GEx	T
Naïve Predictor	Naïve Predictor △	N/A	N/A
PathDNN	PathDNN ★	GEx	T
	PathDNN_rand ◊	GEx	T
CDS	CDS ★	GEx, CNV	T
	CDS_rand ◊	GEx, CNV	T
	CDS (GEx, T) ■	GEx	T
	CDS_rand (GEx, T) ◊	GEx	T
HiDRA	HiDRA ★	GEx	FP
	HiDRA_rand ◊	GEx	FP
	HiDRA (GEx, T) ■	GEx	T
	HiDRA_rand (GEx, T) ◊	GEx	T
PathDSP	PathDSP ★	GEx, CNV, Mut	FP, T
	PathDSP_rand ◊	GEx, CNV, Mut	FP, T
	PathDSP (GEx, FP) ■	GEx	FP
	PathDSP_rand (GEx, FP) ◊	GEx	FP
	PathDSP (GEx, T) ■	GEx	T
	PathDSP_rand (GEx, T) ◊	GEx	T

Each of the pathway-based models used different data modalities in their original study (Supplementary Table S1). We tested all models using the three pathway collections discussed earlier. For implicit models, we tried both drug targets (T) and Morgan fingerprints (FP); however, for explicit models, only drug targets could be used due to their requirements that the drug features must be at gene level. For CCL features, we used all data modalities used by the original study. However, since gene expression data was used by all models (alone or in combination with other omics data, Supplementary Table S1), we also implemented model variants that only utilized GEx data. This ensured that one architecture is not given an unfair advantage due to

234 access to a larger number of modalities. Table 2 provides a summary of all variations of the
235 models considered in this study.
236

237 **Baseline Models**

238 We used three types of baseline models to benchmark the pathway-based models and their
239 variants. First, we used a multilayer perceptron (MLP) with five layers as a universal baseline for
240 all models. This MLP represents a black-box feedforward neural network that is often used for
241 benchmarking of other deep learning architectures (including the pathway-based models). Since
242 all pathway-based models had a variant trained with GEx data, along with drug targets (or
243 Morgan fingerprints), we trained two MLP models, MLP (GEx, FP) and MLP (GEx, T), representing
244 the data input options above (Table 2).
245

246 The second type of baseline used in our study was a predictor that simply calculates the average
247 drug sensitivity measure of samples in the training set and reports their average for all samples
248 in the test set (henceforth referred as naïve predictor). More specifically, the naive predictor does
249 not use any CCL or drug features, but instead simply relies on the identity of the CCL or the drug
250 (depending on the data splitting strategy). As shown in Supplementary Figure S1, in the LCO setup
251 and for a (CCL, drug) pair in the test set, the naive predictor reports the average response of all
252 CCLs in the training set to that drug. As a result, all CCLs in the test set will have the same response
253 value for a drug (i.e., only the drug identity determine the response). On the other hand, in the
254 LDO setup and for a (CCL, drug) pair, the average response of the CCL to all drugs in the training
255 set is reported as the prediction (i.e., only the identity of the CCL determines the response). In

256 the case of LPO, the averaging is done across all drugs and all CCLs corresponding to a (CCL, drug)
257 in the test set. The naïve predictors reveal the performance of a model that does not learn the
258 relationship between the input features and output drug response and can control for inflation
259 in the performance metrics.
260
261 The third type of baselines correspond to model-specific baselines that have the exact same
262 architecture of a pathway-based model (with all their input data), but instead of gene-pathway
263 membership information from pathway databases use randomly generated pathways. This type
264 of baseline model (shown with a suffix of "_rand" in Table 2) allows us to determine if the
265 (potential) performance improvement of a pathway-based model is due to the added value of
266 biological information, or instead is a technical artifact. Let a pathway collection (e.g., KEGG)
267 contain m pathways $P_i, i = 1, 2, \ldots, m$, each with N_i genes. Then, a randomly generated pathway
268 collection was produced by randomly assigning N_i genes to pathway P_i. We evaluated the
269 performance of each pathway-based model with multiple randomly generated pathway
270 collections to determine the mean, standard deviation and histogram of the performance metrics
271 of these random pathway baselines.

Results

Models that incorporate KEGG pathway information implicitly outperformed explicit models

Since KEGG was the most commonly used pathway collection in the original studies (Supplementary Table S1), we used the uniform dataset that we formed for this collection to comprehensively evaluate all models. We first focused on GEx data to represent CCLs since all models used GEx modality in their original studies. We also used drug targets to represent compounds since all models could take advantage of this data modality (Morgan fingerprints are not compatible with PathDNN and CDS). Table 3 shows SCC and RMSE values for LCO, LDO, and LPO data splitting strategies (see Supplementary Table S4 for other performance measures).

Table 3: Performance of pathway-based models using KEGG collection, with GEx and drug targets (T) as inputs. The mean and standard deviations (std) are calculated across CCLs. The best performing model is bold-faced, while worst performing model is underlined. Models are ranked by their LCO RMSE. The following symbols are used in this table: △ = universal baseline, ★ = original pathway-based model, ■ = model variant, ↑ = higher value indicates better performance, ↓ = lower value indicates better performance. ∗ The LDO SCC cannot be calculated for the naive predictor since in this case it outputs the same value for all CCLs.

Model Name	LCO		LDO		LPO	
	SCC ↑ (±std)	RMSE ↓ (±std)	SCC ↑ (±std)	RMSE ↓ (±std)	SCC ↑ (±std)	RMSE ↓ (±std)
PathDSP (GEx, T) ■	**0.882 (±0.045)**	**1.286 (±0.230)**	0.380 (±0.146)	2.648 (±0.287)	**0.876 (±0.074)**	**1.103 (±0.256)**
HiDRA (GEx, T) ■	0.864 (±0.048)	1.368 (±0.253)	0.356 (±0.156)	<u>3.110</u> (±0.400)	0.863 (±0.078)	1.174 (±0.260)
Naïve Predictor △	0.871 (±0.045)	1.373 (±0.292)	NA *	2.826 (±0.274)	0.855 (±0.087)	<u>1.742</u> (±0.280)
MLP (GEx, T) △	0.858 (±0.042)	1.420 (±0.231)	**0.382 (±0.160)**	**2.686 (±0.366)**	0.845 (±0.086)	1.294 (±0.286)
PathDNN (GEx, T)★	0.857 (±0.044)	1.494 (±0.292)	<u>0.342</u> (±0.179)	3.054 (±0.485)	0.851 (±0.083)	1.245 (±0.273)
CDS (GEx, T) ■	<u>0.789</u> (±0.098)	<u>1.603</u> (±0.254)	0.345 (±0.156)	2.724 (±0.317)	<u>0.769</u> (±0.106)	1.508 (±0.304)

PathDSP (GEx, T) outperformed all models in LCO and LPO data splitting schemes, while its performance was second best after the MLP (GEx, T) baseline in the LDO scheme. Compared to the naive predictor, PathDSP (GEx, T) had a better performance in all evaluations, where the highest difference was observed under the LPO validation scheme with 37% lower average RMSE. Overall, the implicit models (HiDRA and PathDSP) outperformed the two universal baselines (MLP and naive predictor) for the majority of evaluations, while the explicit models (PathDNN and CDS) did not outperform them in a considerable number of evaluations (Table 3, Supplementary Tables S3).

Next, we assessed the improvement provided by each method compared to the corresponding naïve predictor (Figure 3A). With regards to RMSE, all models provided improvement for the majority of CCLs in the LPO framework, which is expected since the prediction task in LPO is significantly easier than LCO and LDO. However, in LDO and LCO, many models could not provide a lower RMSE compared to the naïve predictor. The improvement was even less in terms of SCC (Figure 3A). However, PathDSP outperformed the naïve predictor for the majority of CCLs in all data splitting setups in terms of RMSE and SCC.

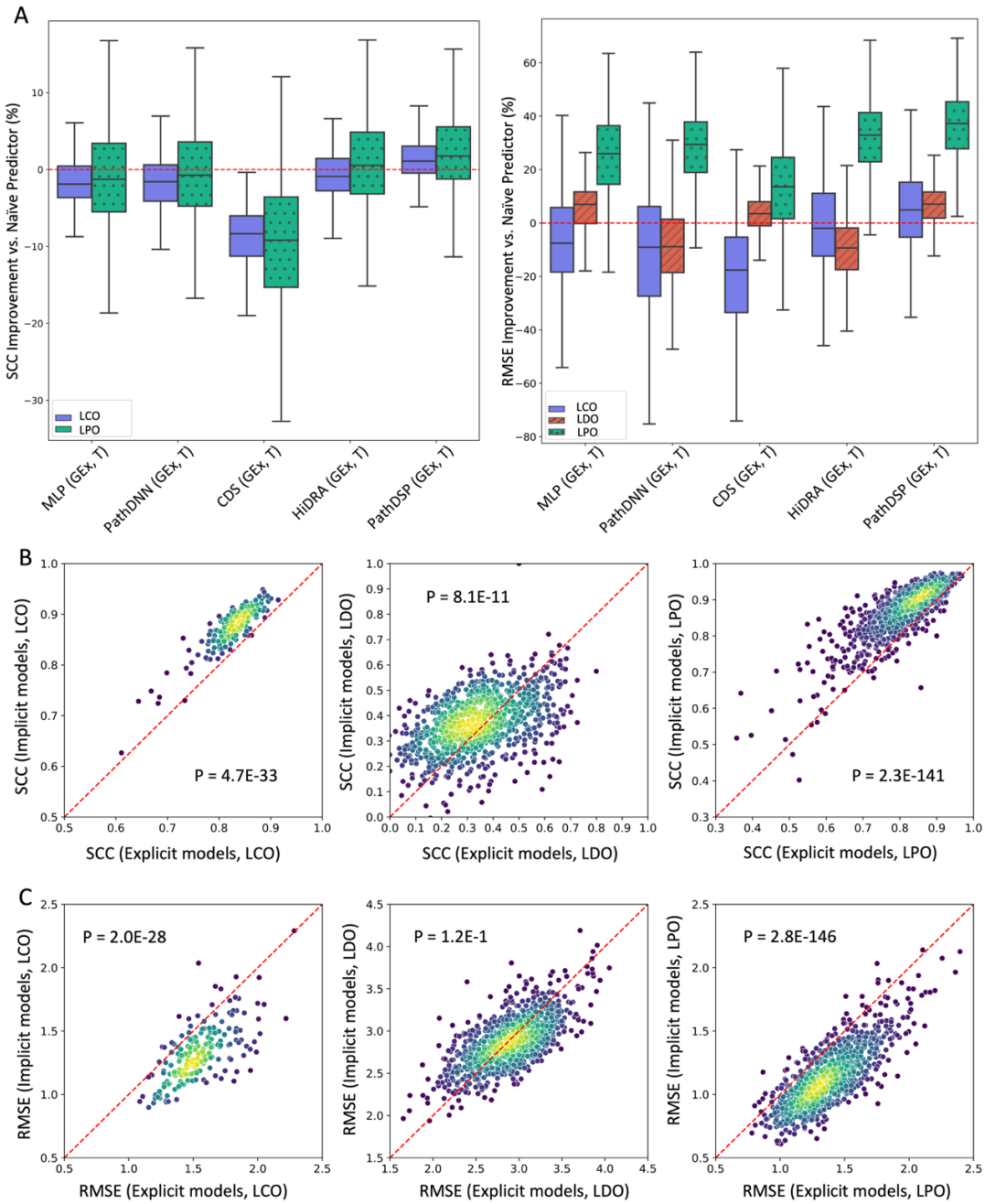

Figure 3: Performance of models in different data splitting setups. A) The improvement of each model versus naïve predictor. Box plots show the distribution of performance improvement for CCLs. Each box shows the range between 25th and 75th percentiles, while whiskers show the minimum and maximum improvement. SCC for naïve predictor cannot be calculated in LDO. B-C) Performance of implicit pathway models versus explicit models that use a pathway layer. Each circle represents a CCL. The color of each circle represents the density of circles in its vicinity. P-values are calculated using a two-sided Wilcoxon signed-rank test. The average performance of explicit models (PathDNN and CDS) is shown on the x-axis, while the performance of implicit models (PathDSP and HiDRA) is shown on the y-axis. Panel B shows the performance in terms of SCC, while panel C shows it in terms of RMSE.

318 Next, we sought to directly compare the performance of explicit models against implicit models.
319 For this purpose, we calculated the average performance of the two implicit (PathDSP (GEx, T),
320 HiDRA (GEx T)) and the two explicit (PathDNN (GEx, T), CDS (GEx, T)) models for each CCL, and
321 used a two-sided Wilcoxon signed rank test to assess if one strategy outperforms the other
322 (Figure 3B). Based on SCC, the implicit strategy significantly outperformed the explicit strategy
323 that utilizes a pathway layer, for all three data splitting strategies. A similar pattern was observed
324 using RMSE, but for LDO strategy the difference was not statistically significant. These results
325 further confirm the observation that utilizing an explicit pathway layer does not seem to perform
326 well in prediction of drug response.
327

328 **Morgan fingerprints of compounds were more informative than drug targets for predicting**
329 **response of unseen cell lines**

330 Since three of the considered models (MLP, HiDRA, and PathDSP) can utilize both drug targets (T)
331 and Morgan fingerprints (FP) to represent drugs, we sought to determine which compound
332 representation is most informative for drug response prediction. As can be seen in Figure 4, in all
333 three models, using FP to represent compounds in most cases was superior in terms of SCC in
334 predicting unseen CCLs (LCO) or in predicting unseen CCL-drug pairs (LPO) (Two-sided Wilcoxon
335 signed-rank $P<0.05$, except for PathDSP LCO). On the other hand, in all three models drug targets
336 were more informative in predicting the response of unseen drugs (LDO). However, one should
337 note that none of the three models performed very well in the LDO data splitting setup and better
338 compound representations may be necessary for such an application.
339

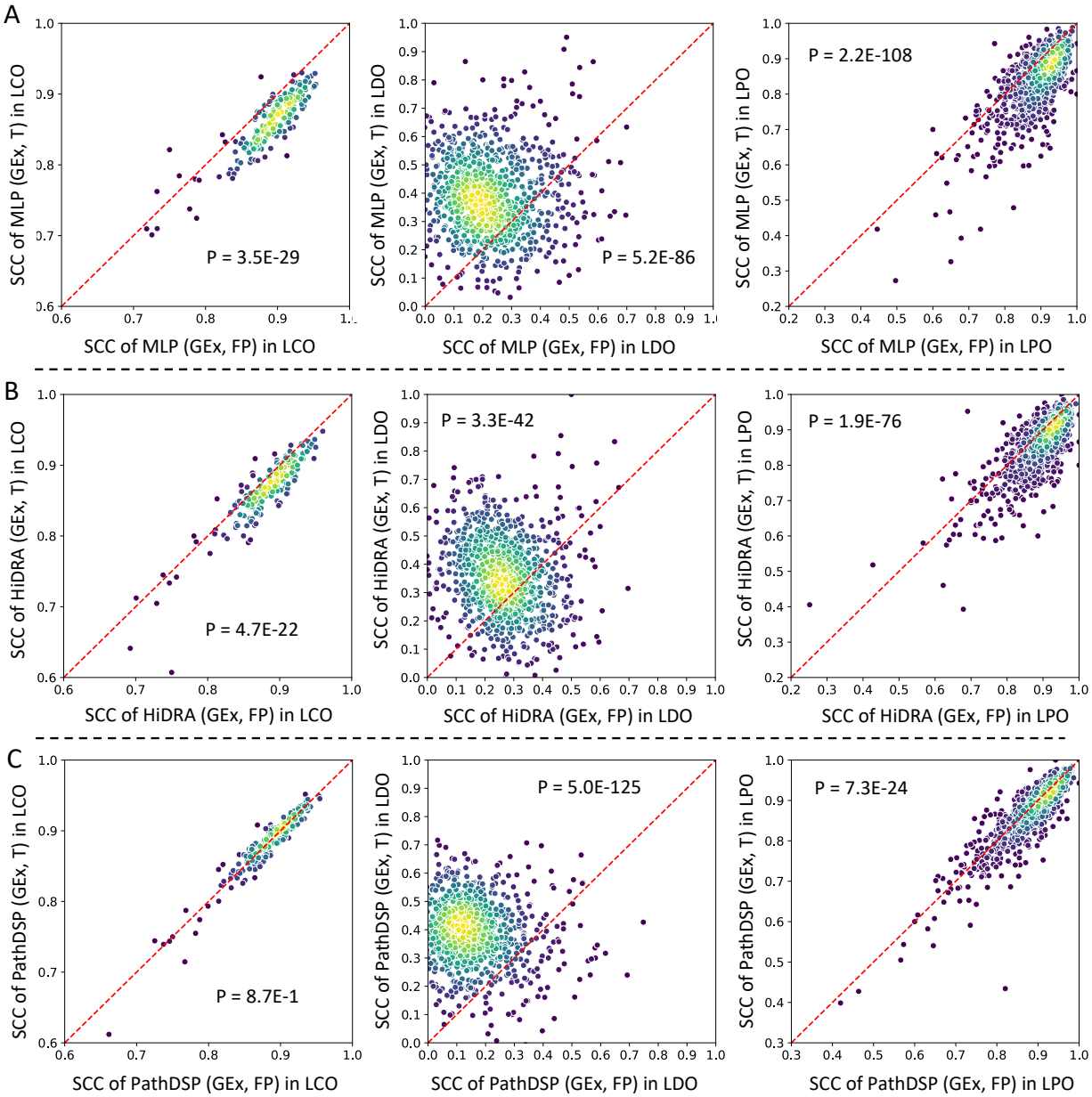

Figure 4: Performance of three models when using drug targets or Morgan fingerprints to represent drugs in terms of SCC. Each circle represents a CCL. The color of each circle represents the density of circles in its vicinity. P-values are calculated using a two-sided Wilcoxon signed-rank test. For all models, the mean SCC when using FP was higher in LCO and LPO, and lower in LDO compared to when using T. A) Performance of MLP (GEx, T) versus MLP (GEx, FP). B) Performance of HiDRA (GEx, T) versus HiDRA (GEx, FP). C) Performance of PathDSP (GEx, T) versus PathDSP (GEx, FP). Only models that could utilize both FP and T to represent drugs were used for this analysis.

It is worth noting that although PathDSP (GEx, T) and HiDRA (GEx, T) both outperformed the MLP (GEx, T) baseline in LCO and LPO evaluation, MLP (GEx, FP) baseline outperformed all other models, independent of which CCL or drug representations they used in both LCO and LPO (Supplementary Table S4). This is an important observation that shows that a simple MLP baseline, when used with appropriate inputs could achieve comparable or better results compared to various interpretable models. This observation is concordant with the observation in [15], where the authors found that the interpretable version of their model resulted in comparable performance to the matched black-box model.

Integrating multiple data modalities improves performance of PathDSP and CDS

Among all methods that we considered in this study, two of them (CDS and PathDSP) used multiple data modalities in their original study (Supplementary Table S1). Table 4 compares the performance of these methods when data modalities chosen by the original study were used as inputs against their performance when only GEx was used.

The original PathDSP model uses GEx, somatic mutation, and CNV as CCL features, as well as Morgan fingerprints and drug targets as compound features. PathDSP outperformed both PathDSP (GEx, T) and PathDSP (GEx, FP) in 5 out of 6 evaluations (all except RMSE in LCO, Table 4). The original CDS model uses GEx and CNV as CCL features and drug targets as compound features. The original CDS also outperforms CDS (GEx, T) in all evaluations except for LCO approach. Overall, these results suggest that using multiple data modalities can improve the performance of each model. However, it is important to note that MLP (GEx, FP) outperformed

all models (including the multi-modality versions of PathDSP and CDS) in LCO and LPO evaluations (Supplementary Table S4).

Table 4: Performance of CDS and PathDSP using KEGG collection, with different input choices. The mean and standard deviations (std) are calculated across CCLs. For each method, the input choice that performs best is bold-faced. Since CDS can only use drug targets to represent compounds, the only considered baseline for it is CDS (GEx, T). The following symbols are used in this table: ★ = original pathway-based model, ■ = model variant, ↑ = higher value indicates better performance, ↓ = lower value indicates better performance.

Model Name	LCO		LDO		LPO	
	SCC ↑ (±std)	RMSE ↓ (±std)	SCC ↑ (±std)	RMSE ↓ (±std)	SCC ↑ (±std)	RMSE ↓ (±std)
PathDSP ★	**0.883** (±0.044)	1.308 (±0.276)	**0.470** (±0.130)	**2.477** (±0.286)	**0.893** (±0.068)	**1.020** (±0.239)
PathDSP (GEx, T) ■	0.882 (±0.045)	1.286 (±0.230)	0.380 (±0.146)	2.648 (±0.287)	0.876 (±0.074)	1.103 (±0.256)
PathDSP (GEx, FP) ■	0.882 (±0.043)	1.297 (±0.227)	0.139 (±0.146)	2.944 (±0.327)	0.887 (±0.068)	1.051 (±0.243)
CDS ★	0.777 (±0.049)	1.625 (±0.189)	**0.378** (±0.164)	**2.606** (±0.320)	**0.776** (±0.083)	**1.478** (±0.287)
CDS (GEx, T) ■	**0.789** (±0.098)	**1.603** (±0.254)	0.345 (±0.156)	2.724 (±0.317)	0.769 (±0.106)	1.508 (±0.304)

Randomly generated pathways provide comparable results to biological pathway collections for prediction of drug response in unseen cell lines

Next, we sought to determine whether the performance of pathway-based models can be attributed to the biological information in the pathways, or if randomly generated pathways can also result in a similar performance. For this purpose, we randomly assigned genes to pseudo-pathways while matching the size of the pathways in the KEGG collection. Figure 5A shows the

390 percentage of improvement in the form of a heat map, where we compared the original pathway-
391 based models (PathDNN, CDS, HiDRA, PathDSP) and their model variants with their
392 corresponding random pathway baselines. Figure 5B and Supplementary Figures S2-S6 show the
393 distribution of SCC and RMSE of different models for each randomly generated pathway
394 collection in different validation schemes.

395

396 As can be seen in Figure 5, in the LCO and LPO evaluations, biological pathways provided almost
397 no improvement compared to their randomly generated counterparts (percentage of
398 improvement between -2% and 2%) for PathDSP, PathDNN, and HiDRA. CDS was the only
399 exception for which an improvement of up to 15% was obtained using biological pathways, which
400 is likely due to its unique architecture that when used with randomly generated pathways results
401 in very sparse connections between its input layer and gene layer. Moreover, it is important to
402 note that despite this improvement, CDS and CDS (GEx, T) had a much worse performance
403 compared to the other models (Figure 5B and Table 3). In the LDO evaluation, PathDNN, CDS,
404 and PathDSP all benefited from the biological pathways (compared to their random counterparts),
405 which may suggest the benefit of incorporating pathway information in predicting the response
406 of unseen drugs.

407

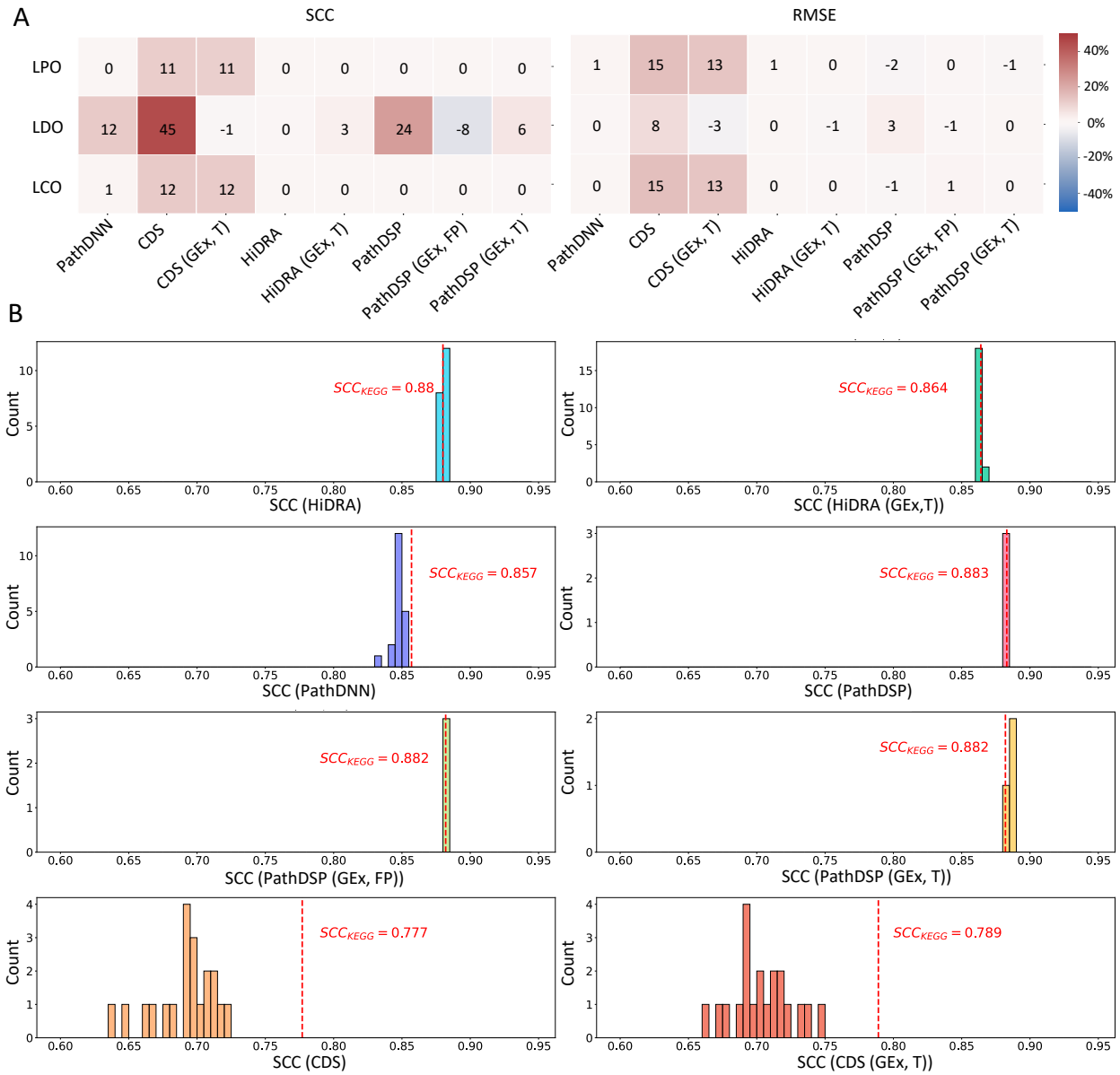

Figure 5: Performance of pathway-based models using KEGG or randomly generated pathways. 1) Percentage of improvement (or deterioration) of different models when using KEGG compared to their mean performance when using randomly generated pathways. B) The histograms show the distribution of mean SCC of random pathway baselines using the LCO validation scheme. Vertical dashed red lines show SCC of the model when using KEGG pathways. Twenty random pathway baselines were constructed for each model, except for PathDSP. Since PathDSP requires 1000 permutation tests for each type of input data, only three random pathway baselines were constructed due to its extremely high computational requirement.

417 **Effect of pathway collection choice on drug response prediction**

418 We next sought to investigate which pathway collection is more suitable for the drug response
419 prediction task. For this purpose, we compared the performance of each pathway-based model
420 (in their original architecture and using original input features) using each of these collections
421 (see Supplementary Tables S5 and S6 for the performance of all models and their variants using
422 PID and Reactome). To ensure a fair comparison, we only included (drug, CCL) pairs in the test
423 sets that were shared among all three uniform datasets. We used the LCO data splitting approach,
424 since the overlap among the test samples of the three uniform datasets was largest in this
425 strategy (21525 pairs versus 3277 in LDO and 851 in LPO).

426

427 **Table 5:** Performance of pathway-based models using different pathway collections. Mean and
428 standard deviation are calculated across cell lines using the LCO evaluation.

Pathway Collection	**PathDNN**		**CDS**		**HiDRA**		**PathDSP**	
	SCC (±std)	RMSE (±std)	SCC (±std)	RMSE (±std)	SCC (±std)	RMSE (±std)	SCC (±std)	RMSE (±std)
Reactome	0. 86 (±0.04)	1.35 (±0.24)	0.76 (±0.05)	1.63 (±0.24)	0.88 (±0.05)	1.26 (±0.25)	0.87 (±0.05)	1.29 (±0.25)
PID	0.85 (±0.05)	1.42 (±0.26)	0.78 (±0.06)	1.63 (±0.27)	0.88 (±0.05)	1.28 (±0.29)	0.88 (±0.05)	1.29 (±0.25)
KEGG	0.85 (±0.05)	1.49 (±0.29)	0.77 (±0.05)	1.61 (±0.20)	0.87 (±0.05)	1.29 (±0.26)	0.88 (±0.05)	1.30 (±0.28)

429

430 Table 5 shows the mean and standard deviation of SCC and RMSE of each model using all three
431 pathway collections. Overall, we observed that the performance of most models did not vary
432 drastically based on the choice of pathway collection. However, Reactome pathway provided
433 slightly better results for the majority of the methods, being the top performing option for 3 (out

434 of 4 methods) based on RMSE. This is likely due to the larger number of pathway annotations
435 included in this database for our use-case (1608 pathways) compared to KEGG (332 pathways)
436 and PID (196 pathways), resulting in a more comprehensive representation of the input data.

437

438 **Discussion**

439 Recently, several deep learning methods have been proposed to enable a higher interpretability
440 of drug response prediction and to improve the prediction performance. In this study, we set out
441 to investigate four methods that try to achieve these goals by incorporating pathway information
442 under various validation settings and answer five important questions discussed earlier. The
443 models were tested to predict drug response for unseen (CCL, drug) pairs, unseen CCLs, and
444 unseen drugs. We compared these methods against three types of baseline models, two of which
445 were usually overlooked in previous studies.

446

447 First, we observed that models that incorporate a dedicated explicit pathway layer and connect
448 genes in the input layer to pathways based on pathway membership perform worse compared
449 to models that implicitly (e.g., using attention mechanisms or pathway enrichment scores)
450 incorporate pathway information. In fact, on many occasions explicit models' performance was
451 inferior to a black-box simple MLP model with similar input. This suggests that direct encoding of
452 gene-pathway membership is not an effective strategy to incorporate pathway information. The
453 overly sparse connections between the gene and pathway layer may be the cause for the
454 unsatisfactory performance of these methods. Another limitation of explicit models is that they
455 can only utilize gene-level drug representations, limiting them to drug targets. Our analysis using

456 methods that could utilize both drug targets and Morgan fingerprints showed the latter to be
457 superior in prediction of response for unseen CCLs or unseen pairs.
458
459 Our analyses also showed that while implicit models generally perform better in predicting
460 unseen CCLs and unseen pairs, a comparable performance can be achieved when instead of
461 biological pathways, randomly generated pathways are used. Moreover, the best performance
462 in these validation setups were achieved using a black-box MLP that used Morgan fingerprints to
463 represent drugs. Put together, these results suggest that to make the models interpretable, these
464 approaches inevitably make assumptions that cannot fully capture the nuances of drugs'
465 mechanisms of action in cancer cell lines, resulting in comparable or worse performance to black-
466 box models, an observation also made by Kuenzi et al. [15].
467
468 Our analyses also allowed us to assess the difficulty of drug response prediction in different
469 setups. While at first glance Table 3 may suggest that predicting response of unseen drugs are
470 much more challenging than unseen CCLs, a more appropriate comparison can be made using
471 Figure 3A, in which the performance improvement compared to a naïve predictor is shown. This
472 figure shows that predicting response of unseen pairs is much easier compared to prediction for
473 unseen CCLs and unseen drugs. This is not surprising, since this is a transductive setup and drugs
474 and CCLs in the test set are present in the training set (but not together). This setup is useful for
475 imputation of missing drug response values but cannot be used to predict response to new CCLs
476 or new drugs. On the other hand, predicting response of unseen drugs and unseen CCLs are much
477 more difficult and most models cannot provide a better prediction for the majority of CCLs in

478 these two setups (Figure 3A). The contrast between conclusions one may draw from Table 3 and
479 Figure 3A demonstrates the potential for obtaining inflated performance measures in the LCO
480 framework. Since the distribution (across CCLs) of log IC50 values of different drugs may be
481 drastically different from each other, by simply knowing the identity of the drug, a model can
482 rank different drugs based on their log IC50 values for an unseen CCL. To overcome this issue and
483 avoid reporting unrealistically inflated metrics, one should focus on improvement compared to a
484 naïve predictor (an approach that we adopted in this study), or should normalize log IC50 values
485 of each drug across CCLs to make them comparable to each other (an approach that we adopted
486 in [4]).
487
488 We also compared the performance of different models using three pathway collections, PID,
489 KEGG, and Reactome. Even though there was not a major difference in the performance of
490 models when substituting one collection for the other, Reactome collection resulted in slightly
491 better performance. This may be due to the larger number of pathways in this collection
492 compared to the other two. However, since randomly generated pathway collections also
493 provided comparable performance based on the models considered in this study, it is not
494 possible to draw a conclusive determination regarding which pathway collection may be more
495 useful for the drug response prediction task.
496
497 This study focused on evaluating the effect of incorporating pathway information from the
498 perspective of model performance and we did not evaluate these models based on their level of
499 interpretability. A study that focuses on interpretability aspect of these models would be very

500 insightful and complementary to the current study. For example, one can take a closer look at
501 the feature attributions of these pathway-based models using explainers such as DeepLIFT (Deep
502 Learning Important FeaTures) [30], CXPlain [31], and SHAP (Shapley Additive exPlanations) [32]
503 to estimate feature importance and identify genes or other biological features that have
504 substantial influence on the model predictions. Such analysis can be done for all pathway-based
505 models to check if the most important/predictive sub-networks or the top contributing genes
506 extracted from each model have any overlap. If such overlap exists between the pathway-based
507 models, further studies can be done by validating the findings with existing literature or
508 conducting experiments under a lab setting. Analysis on model interpretability will complement
509 the insights obtained from model performance evaluation and together provide a more holistic
510 view for the effect of pathway incorporation on drug response prediction.
511
512 In conclusion, we believe that while interpretability is a very crucial aim in precision medicine,
513 new models are necessary to enable a higher degree of interpretability while at the same time
514 improve the drug response prediction performance. In addition, it is not sufficient for these
515 models to show a better performance compared to their black-box counterparts, and they need
516 to also evaluate their models against randomly generated pathways (with similar pathway sizes
517 to the original collection) and naïve predictors to control for different types of biases.
518
519 **Data and Code Availability**
520 All data generated as part of this study are provided as supplementary files. The input data for
521 the evaluated models is provided at https://zenodo.org/record/7101665#.YzS79HbMKUk. The

link https://github.com/Emad-COMBINE-lab/InterpretableAI_for_DRP contains the code for all the methods evaluated in this study.

Funding:

This work was supported by the Government of Canada's New Frontiers in Research Fund (NFRF) [NFRFE-2019-01290] (AE) and by Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2019-04460 (AE). This work was also funded by Génome Québec, the Ministère de l'Économie et de l'Innovation du Québec, IVADO, the Canada First Research Excellence Fund and Oncopole, which receives funding from Merck Canada Inc. and the Fonds de Recherche du Québec – Santé (AE). This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

Competing Interets

None of the authors have any competing interests.

537 **References**

538 1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G,
539 Thrun S, Dean J: **A guide to deep learning in healthcare.** In *Nature Medicine*, vol. 25. pp.
540 24-29; 2019:24-29.

541 2. Rajkomar A, Dean J, Kohane I: **Machine Learning in Medicine.** *New England Journal of*
542 *Medicine* 2019, **380:**1347-1358.

543 3. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Ammad-Ud-
544 Din M, Hintsanen P, Khan SA, et al: **A community effort to assess and improve drug**
545 **sensitivity prediction algorithms.** *Nature Biotechnology* 2014, **32:**1202-1212.

546 4. Hostallero DE, Li Y, Emad A: **Looking at the BiG picture: incorporating bipartite graphs in**
547 **drug response prediction.** *Bioinformatics* 2022, **38:**3609-3620.

548 5. El Khili MR, Memon SA, Emad A: **MARSY: A multitask deep learning framework for**
549 **prediction of drug combination synergy scores.** *bioRxiv* 2022**:**2022.2006.2007.495155.

550 6. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu FX, Wang J: **DeepDSC: A Deep Learning Method to**
551 **Predict Drug Sensitivity of Cancer Cell Lines.** *IEEE/ACM Transactions on Computational*
552 *Biology and Bioinformatics* 2021, **18:**575-582.

553 7. Rampášek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A: **Dr.VAE: Improving drug**
554 **response prediction via modeling of drug perturbation effects.** *Bioinformatics* 2019,
555 **35:**3743-3751.

556 8. Malioutov DM, Varshney KR, Emad A, Dash S: **Learning Interpretable Classification Rules**
557 **with Boolean Compressed Sensing.** In *Transparent Data Mining for Big and Small Data.*

558 Edited by Cerquitelli T, Quercia D, Pasquale F. Cham: Springer International Publishing;
559 2017: 95-121
560 9. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S,
561 Gil-Lopez S, Molina D, Benjamins R, et al: **Explainable Artificial Intelligence (XAI):**
562 **Concepts, taxonomies, opportunities and challenges toward responsible AI.** *Information*
563 *Fusion* 2020, **58:**82-115.
564 10. Azodi CB, Tang J, Shiu SH: **Opening the Black Box: Interpretable Machine Learning for**
565 **Geneticists.** *Trends Genet* 2020, **36:**442-455.
566 11. Hostallero DE, Wei L, Wang L, Cairns J, Emad A: **A Deep Learning Framework for**
567 **Prediction of Clinical Drug Response of Cancer Patients 1 and Identification of Drug**
568 **Sensitivity Biomarkers using Preclinical Samples 2 3.** *bioRxiv*
569 2021**:**2021.2007.2006.451273.
570 12. Baptista D, Correia J, Pereira B, Rocha M: **Evaluating molecular representations in**
571 **machine learning models for drug response prediction and interpretability.** *Journal of*
572 *Integrative Bioinformatics* 2022.
573 13. Che Z, Purushotham S, Khemani R, Liu Y: **Interpretable Deep Models for ICU Outcome**
574 **Prediction.** *AMIA Annu Symp Proc* 2016, **2016:**371-380.
575 14. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N: **Intelligible Models for**
576 **HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission.** In
577 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery*
578 *and Data Mining*. pp. 1721–1730. Sydney, NSW, Australia: Association for Computing
579 Machinery; 2015:1721–1730.

580 15. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, Ma J, Ideker T: **Predicting Drug**
581 **Response and Synergy Using a Deep Learning Model of Human Cancer Cells.** *Cancer Cell*
582 2020, **38:**672-684.e676.
583 16. Deng L, Cai Y, Zhang W, Yang W, Gao B, Liu H: **Pathway-guided deep neural network**
584 **toward interpretable and predictive modeling of drug sensitivity.** *Journal of Chemical*
585 *Information and Modeling* 2020, **60:**4497-4505.
586 17. Jin I, Nam H: **HiDRA: Hierarchical Network for Drug Response Prediction with Attention.**
587 In *Journal of Chemical Information and Modeling*, vol. 61. pp. 3858-3867; 2021:3858-3867.
588 18. Snow O, Sharifi-Noghabi H, Lu J, Zolotareva O, Lee M, Ester M: **Interpretable Drug**
589 **Response Prediction using a Knowledge-based Neural Network.** In *Proceedings of the*
590 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp.
591 3558-3568; 2021:3558-3568.
592 19. Tang YC, Gottlieb A: **Explainable drug sensitivity prediction through cancer pathway**
593 **enrichment.** *Scientific Reports* 2021, **11**.
594 20. Zhang H, Chen Y, Li F: **Predicting Anticancer Drug Response With Deep Learning**
595 **Constrained by Signaling Pathways.** *Frontiers in Bioinformatics* 2021, **1**.
596 21. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** In *Nucleic Acids*
597 *Research*, vol. 28. pp. 27-30; 2000:27-30.
598 22. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: The**
599 **pathway interaction database.** *Nucleic Acids Research* 2009, **37**.

600 23. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, D'Eustachio P,
601 Stein L, Hermjakob H: **Reactome pathway analysis: A high-performance in-memory**
602 **approach.** *BMC Bioinformatics* 2017, **18**.
603 24. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith
604 JA, Thompson IR, et al: **Genomics of Drug Sensitivity in Cancer (GDSC): A resource for**
605 **therapeutic biomarker discovery in cancer cells.** *Nucleic Acids Research* 2013, **41**.
606 25. Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M: **STITCH 5: Augmenting**
607 **protein-chemical interaction networks with tissue and affinity data.** *Nucleic Acids*
608 *Research* 2016, **44:**D380-D384.
609 26. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et
610 al: **PubChem in 2021: New data content and improved web interfaces.** *Nucleic Acids*
611 *Research* 2021, **49:**D1388-D1395.
612 27. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva
613 NT, Morris JH, Bork P, et al: **STRING v11: Protein-protein association networks with**
614 **increased coverage, supporting functional discovery in genome-wide experimental**
615 **datasets.** *Nucleic Acids Research* 2019, **47:**D607-D613.
616 28. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D: **InChI, the IUPAC International**
617 **Chemical Identifier.** *Journal of Cheminformatics* 2015, **7**.
618 29. Landrum G: **RDKit: Open-source Cheminformatics.** *Http://WwwRdkitOrg/* 2006.
619 30. Shrikumar A, Greenside P, Kundaje A: **Learning important features through propagating**
620 **activation differences.** In *34th International Conference on Machine Learning, ICML 2017*,
621 vol. 7. pp. 4844-4866; 2017:4844-4866.

31. Schwab P, Karlen W: **CXPlain: Causal explanations for model interpretation under uncertainty.** In *Advances in Neural Information Processing Systems*, vol. 32; 2019.

32. Lundberg SM, Lee SI: **A unified approach to interpreting model predictions.** In *Advances in Neural Information Processing Systems*, vol. 2017-December. pp. 4766-4775; 2017:4766-4775.