Molecular Omics

ROYAL SOCIETY OF CHEMISTRY

RESEARCH ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Mol. Omics*, 2021, **17**, 108

DeepSIBA: chemical structure-based inference of biological alterations using deep learning†

C. Fotis, ‡ N. Meimetis, ‡ A. Sardis and L. G. Alexopoulos*

Received 23rd September 2020,
Accepted 9th November 2020

DOI: 10.1039/d0mo00129e

rsc.li/molomics

Predicting whether a chemical structure leads to a desired or adverse biological effect can have a significant impact for *in silico* drug discovery. In this study, we developed a deep learning model where compound structures are represented as graphs and then linked to their biological footprint. To make this complex problem computationally tractable, compound differences were mapped to biological effect alterations using Siamese Graph Convolutional Neural Networks. The proposed model was able to encode molecular graph pairs and identify structurally dissimilar compounds that affect similar biological processes with high precision. Additionally, by utilizing deep ensembles to estimate uncertainty, we were able to provide reliable and accurate predictions for chemical structures that are very different from the ones used during training. Finally, we present a novel inference approach, where the trained models are used to estimate the signaling pathway signature of a compound perturbation, using only its chemical structure as input, and subsequently identify which substructures influenced the predicted pathways. As a use case, this approach was used to infer important substructures and affected signaling pathways of FDA-approved anticancer drugs.

1. Introduction

Early stage drug discovery aims to identify the right compound for the right target, for the right disease. A very important step in this process is hit identification, in which compounds that exhibit strong binding affinity to the target protein are prioritized. Traditionally, the most widely employed method for *in vitro* hit identification is High Throughput Screening (HTS). *In vitro* HTS can produce hits with strong binding affinity that may later be developed into lead compounds through lead optimization. However, due to the vast chemical space, even large scale *in vitro* HTS offers limited chemical coverage. On this front, the development of Computer Aided Drug Design (CADD) methods has enabled the virtual High Throughput Screening (vHTS) of vast compound libraries, thus effectively increasing the search space of hit identification. CADD methods for vHTS focus on compounds' chemical structures and prioritize those that are likely to have activity against the target, for further experiments.[1] More specifically, ligand-based approaches are based on the hypothesis that similar chemical structures will cause similar biological response, by binding to the same protein.[2] However, there are many cases of compounds and drugs, which although structurally dissimilar, cause similar biological effect, either because of off-target effects or by targeting proteins in the same pathway.[3] As a whole, CADD approaches focus on optimal binding affinity, by assessing a compound's structural attributes, often disregarding the effect of the perturbation on the biological system, which is closely related to clinical efficacy and toxicity.[4]

Advances in systems-based approaches and 'omics technologies have led to the development of systems pharmacology methods that aim to lower the attrition rates of early stage drug discovery. Systems pharmacology approaches couple 'omics data with knowledge bases of molecular interactions and network analysis methods in order to assess compounds based on their biological effect.[5] One approach that has gained considerable attraction is the use of gene expression (GEx) profiling to characterize the systematic effects of compounds. On this front, Verbist *et al.* showed how GEx data were able to influence decision making in eight drug discovery projects by uncovering potential adverse effects of the lead compounds.[6] Additionally, Iorio *et al.* utilized similarities between drugs' transcriptional responses to create a drug network and identified the mechanism of action of new drugs based on their position in the network.[7] Since its release, the Connectivity Map (CMap) and the LINCS project have been a cornerstone of transcriptomic-based approaches by providing a large scale database of transcriptomic signatures from compound perturbations along with essential signature matching algorithms.[8,9] CMap's approach is based on the hypothesis that compounds

Biomedical Systems Laboratory, National Technical University of Athens, Athens, Greece. E-mail: leo@mail.ntua.gr

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0mo00129e

‡ Equal contributions.

with similar transcriptomic signatures will cause similar physiological effects on the cell and has been widely adopted by the field of drug repurposing.[10] However, signature-based approaches are not only limited in the search space of compounds with available GEx data but are also missing key structural information that is pivotal for drug design. Thus, an interdisciplinary framework that translates a compound's structural attributes to its biological effect holds promise in augmenting the application of both CADD and systems-based approaches for drug discovery. A computational approach that meets the requirements of such an interdisciplinary framework is Machine Learning (ML) and especially Deep Learning (DL).

The recent increase in available data and computing power has given rise to Deep Learning (DL) methods for various drug discovery tasks, including bioactivity and toxicity prediction as well as *de novo* molecular design.[11–15] DL methods offer the advantage of flexible end-to-end architectures that learn task specific representations of chemical structures, without the need for precomputed features.[16] One particular DL architecture that has achieved state of the art results in several drug discovery benchmark datasets is the Graph Convolutional Neural Network (GCNN).[17,18] Molecular GCNNs operate on chemical structures represented as undirected graphs, with nodes being the atoms and edges the bonds between them. Kearnes *et al.* developed the Weave graph convolution module, which encodes both atom and bond representations and combines them using fuzzy histograms to extract meaningful molecule-level representations.[19] Despite their improved performance over traditional ML methods, end-to-end models including GCNNs are still prone to generalization errors on new chemical scaffolds. This is mainly because of the limited coverage of the chemical space by the training data.[20] In order to tackle this limited chemical coverage, methods like one-shot learning are promising candidates for drug discovery applications. One-shot learning techniques, such as Siamese networks, aim to learn a meaningful distance function between related inputs and have shown increased performance over traditional methods in tasks with few data points.[21–24] Altae-Tran *et al.* implemented one-shot learning for drug discovery by combining graph convolutions and Long Short Term Memory (LSTM) networks with attention and achieved better results than traditional GCNNs.[25] Furthermore, for drug discovery applications, uncertainty estimation is crucial, since incorrect predictions *e.g.* regarding toxicity can lead to incorrect prioritization of compounds for further experimental testing.[26–29] On this front, Ryu *et al.* developed Bayesian GCNNs for molecular property, bioactivity and toxicity predictions and showed that quantifying predictive uncertainty can lead to more accurate virtual screening results.[30] The flexibility provided by GCNN architectures along with one-shot learning and uncertainty estimation approaches can combine aspects from both systems and ligand-based methods into an interdisciplinary framework for early stage drug discovery.

In this paper, we employ deep learning to decipher the complex relationship between a compound's chemical structure and its biological effect. To make this complex problem computationally tractable, we focus on learning a combined representation and distance function that maps structural differences to biological effect alterations. For this task, we propose a deep Siamese GCNN model called deepSIBA. DeepSIBA takes as input pairs of compound structures, represented as graphs and outputs their biological effect distance, in terms of enriched biological processes (BPs) along with an estimated uncertainty. DeepSIBA is trained to minimize the loss between predicted and calculated distances of enriched BPs for compound pairs with available GEx data. In order to account for the biological factors that influence the learning task, we train cell line-specific deep ensembles only on carefully selected chemical structures, for which high quality GEx data are available. The performance of our approach was evaluated with a realistic drug discovery scenario in mind, where gene expression data are available for only one compound per pair and compared with ML methods for pairwise (dyadic) data.[31,32] Finally, we present a novel inference approach, in which the trained models can be used to infer the signaling pathway signature of a target compound, without available GEx data. This inference approach is coupled with a novel method, based on graph saliency maps,[33] which can identify substructures that are responsible for a compound's inferred biological footprint. As a use case, this approach was tasked to infer the signaling pathway signature and important substructures of approved anticancer drugs for which no transcriptomic signatures are available in our data sets, using only their chemical structure as input. DeepSIBA can be used in combination with existing *in silico* drug discovery pipelines to identify structures that not only exhibit maximal binding affinity but also cause a desired biological effect. Thus, by incorporating deepSIBA's interdisciplinary approach, the drug discovery process can produce candidates with improved clinical efficacy and toxicity.

2. Material and methods

2.1 The deepSIBA approach

The overview of our approach is presented in Fig. 1. Transcriptomic signatures from compound perturbations along with their respective chemical structures were retrieved from the CMap dataset.[9] For each compound perturbation, normalized enrichment scores (NES) of GO terms related to BPs were calculated using Gene Set Enrichment Analysis (GSEA). Afterwards, the lists of enriched BPs were ranked based on NES and a Kolmogorov–Smirnov based distance function, similar to GSEA, was used to calculate their pairwise distance (Fig. 1A). During the learning phase, the proposed model is trained to predict the pairwise distance between compounds' affected BPs using only their chemical structure as input. The input chemical structures are represented as undirected graphs, with nodes being the atoms and edges the bonds between them and encoded using a Siamese GCNN architecture (Fig. 1B). In our approach, compounds with available GEx data, representing a small portion of the chemical space, serve as reference for the inference phase. During inference, the model is tasked to

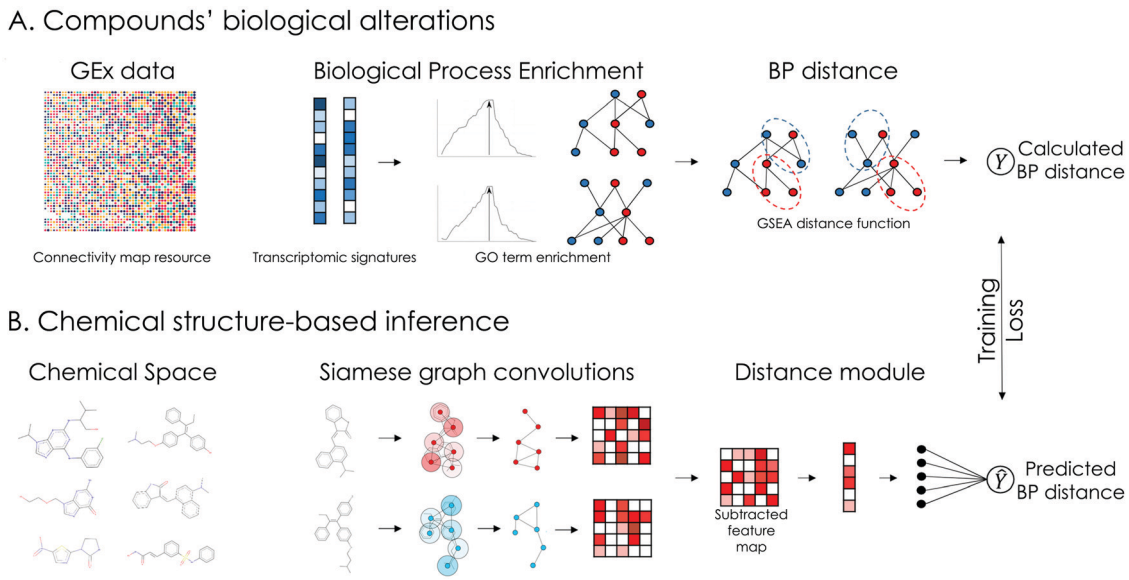

Fig. 1 Schematic overview of deepSIBA. (A) Pairs of transcriptomic signatures following compound treatment are retrieved and enriched GO terms for BPs are calculated. The pairwise distance between enriched BPs is calculated using a Kolmogorov–Smirnov based function (Y). (B) Pairs of chemical structures are represented as molecular graphs and encoded by a deep learning model using Siamese graph convolutions. Compounds' feature maps are then subtracted and a score, which represents their distance between enriched BPs, is predicted ($\hat{Y}$). The deep learning model is trained by minimizing the loss between predicted ($\hat{Y}$) and calculated distance (Y).

predict the biological effect distance between reference and unknown compounds (without available GEx data).

2.2 Data preprocessing and quality control

Transcriptomic signatures (level 5 *z*-score transformed) following compound treatment were downloaded from the L1000 CMap resource.[34] In this project, only the differential expression of the 978 landmark genes in the L1000 assay was considered. For each signature, a quality score was derived, based on its transcriptional activity score (TAS), the number of biological replicates and whether the signature is considered an exemplar. This quality score ranges from Q1 to Q8, with Q1 representing the highest quality. TAS is a metric that measures a signature's strength and reproducibility and is calculated as the geometric mean of the number of differentially expressed (DEx) transcripts and the 75th quantile of pairwise replicate correlations. Furthermore, exemplar signatures are specifically designated for further analysis in the CLUE platform.[35] For each compound per cell line, among signatures from different dosages and time points, the signature with the highest quality was selected. An overview of the processed dataset is presented in ESI† 1.1.

2.3 Biological process enrichment and pairwise distance calculation

Gene ontology (GO) terms for biological processes (BP) involving the landmark genes of the L1000 assay were retrieved using the topGO R package in Bioconductor.[36] Only GO terms with at least 10 genes were considered. For each signature, GO term enrichment was calculated using the R package FGSEA in Bioconductor.[37] Thus, the gene-level feature vector of each perturbation was transformed to a BP-level feature vector of Normalized Enrichment Scores (NES). Pairwise distances between BP-level feature vectors were calculated similar to Iorio *et al.*,[7] using the R package Gene Expression Signature in Bioconductor.[38] Given two feature vectors ranked by NES, A and B, GSEA is used to calculate the ES of the top and bottom GO terms of A in B and *vice versa*. The distance between the vectors is computed as $1 - \frac{ES_{A\text{ in }B} + ES_{B\text{ in }A}}{2}$ and ranges from 0 to 2. An important parameter that can introduce bias in the distance calculation is the number of top and bottom GO terms to consider during GSEA. On this front, an ensemble approach was developed, by calculating pairwise distances between BP-level feature vectors for 5 different numbers of top and bottom GO terms. The numbers we considered were selected based on the average number of significantly enriched GO terms across all perturbations in the dataset (see ESI† 1.3 for details). The distance scores were finally averaged and normalized between 0 and 1.

2.4 Siamese GCNN architecture

A schematic representation of our model's architecture is presented in Fig. 2. The learning model takes as input the chemical structures of compound pairs and predicts their biological distance, at the level of affected biological processes (GO terms). Regarding the input, chemical structures are represented as undirected graphs, with nodes being the atoms and edges the bonds between them. Each input is encoded using 3 matrices: the atom array, which contains atom-level features, the bond array, which contains bond-level features and the edge array, which describes the connectivity of the compound (see ESI† 2.1 for details). The learning model

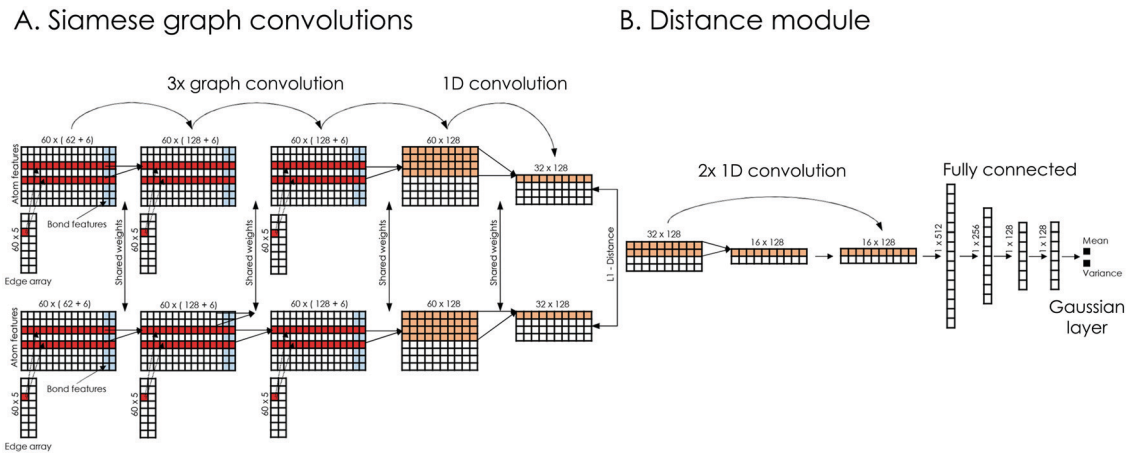

Fig. 2 Schematic representation of the model's architecture. (A) Siamese graph convolutional encoders; compounds' molecular graphs are encoded using 2 encoders with shared weights (Siamese). Each encoder consists of 3 graph convolution and 1 convolution layers. (B) Architecture of the distance module; the distance module consists of 2 convolution, 3 fully connected and 1 Gaussian regression layers.

consists of two Siamese encoders (shared weights) that embed the input graphs into a high dimensional latent space and a trainable distance module that outputs the final distance prediction. The Siamese encoders consist of 3 graph convolutional layers that learn neighborhood-level representations, followed by a convolutional layer that extracts compound-level features (Fig. 2A). Graph convolutions were implemented similar to Duvenaud *et al.*[17] (see ESI† 2.2 for details). The overall goal of the Siamese encoder is to learn task-specific compound representations. The feature maps of the last Siamese layer are then subtracted and their absolute difference is passed to the distance module. The distance module consists of 2 convolutional layers, which extract important features from the difference of the feature maps and 3 fully connected layers that aim to combine those features, while progressively reducing the dimensions (Fig. 2B). Finally, a Gaussian regression layer outputs a mean and variance of the biological effect distance between the compound pair. By treating the distance as a sample from a Gaussian distribution with the predicted mean and variance, the model is trained end-to-end by minimizing the negative log-likelihood criterion[27] given by

$$-\log p_\theta(y_n|X_n) = -\frac{1}{2}\log \sigma_\theta^2(x) - \frac{1}{2\sigma_\theta^2(x)}(y-\mu_\theta(x))^2 + \text{constant}.$$

For each cell line, an ensemble model combining 50 models was created. The ensemble's output is also a Gaussian, with mean and variance calculated from the uniformly weighted mixture of each model. The coefficient of variation (CV) of the Gaussian mixture is used as the model's estimate of predictive uncertainty. The model's hyperparameters, along with the equations for the Gaussian mixture's mean and variance are presented in ESI† 2.3 and 2.4.

2.5 Dataset splitting and evaluation metrics

For each cell line, available compounds were split into training and test. Each cell line specific training set consists of the pairwise distances between training compounds' affected BPs, while each test set contains distances between test and training compounds. Additionally, the Tanimoto similarity between the ECFP4 fingerprints of all training and test compounds was calculated and test compounds that exhibited a similarity higher than 0.85 to any training compound were excluded. An overview of the training and test sets is presented in ESI† 4.1. Across all test scenarios, model performance was evaluated in terms of Mean Squared Error (MSE), Pearson's r and precision. MSE and Pearson's r were calculated between the predicted and computed distance values. In order to calculate precision, the continuous distance values were transformed to binary form by comparing them with an appropriate distance threshold. Even though the learning task is a regression problem, given its nature and potential applications, high precision $\left(\frac{\text{true positives}}{\text{positives}}\right)$ is important in order to avoid false positive hits for validation experiments. The appropriate distance threshold for precision was set at 0.2, based on the distance distribution of duplicate compound signatures, the threshold equivalent to a 90% Connectivity Score and the relationship between the threshold and the actual average number of common enriched BPs. Duplicate signatures indicate transcriptomic signatures from the same compound perturbation, cell line, dose and time point that were assayed on different L1000 plates. Thus, the distribution of distances between duplicate signatures most closely approximates the reference distribution of truly similar biological effect. A thorough investigation of the distance threshold to distinguish compounds with similar biological effect is presented in ESI† 5.1.

2.6 Signaling pathway inference for target structure

The predictions of a trained deepSIBA model can be used to infer a pathway signature for a target structure without the need for GEx data, in terms of the most upregulated and downregulated signaling pathways. The inference approach is similar to the *k*-Nearest Neighbor algorithm (KNN). Given a target structure, a trained ensemble model for the cell line of choice

is used to predict all pairwise distances between target and training compounds. The predicted distance represents the difference between compounds' enriched BPs (GO terms). Training set compounds with predicted distance less than a specified threshold d_{th} are selected as the target's neighbors. If a target structure has more than k neighbors, a signaling pathway signature can be inferred in the following way. For each neighbor N_i, the lists of the top 10 most upregulated and most downregulated pathways, based on NES, are constructed. Pathway enrichment is calculated using FGSEA with KEGG as a knowledge base.[39] KEGG signaling pathways were chosen for inference due to their interpretability. Signaling pathways that appear in the neighbors' lists with a frequency score higher than a threshold f_{th} are selected. Additionally, to account for signaling pathways that are frequently upregulated or downregulated in the set of training compounds, a p-value for each inferred pathway is also calculated. On this front, sets of k neighbors are randomly sampled 5000 times from the training set and a Null distribution of frequency scores for each pathway is derived. A p-value is computed as the sum of the probabilities of observing equally high or higher frequency scores. Finally, only pathways with p-value lower than a threshold p_{th} are inferred. Thus, for each chemical structure, our approach infers two signatures of variable length (up to 10 each) of potentially downregulated and upregulated pathways respectively. For the MCF7 cell line, the aforementioned thresholds and parameters of the inference approach were selected by evaluating the results, in terms of precision and number of inferred pathways, on its respective test set (see ESI† 6.1 for details).

2.7 Substructure importance using graph-based gradients

A graph-based gradient approach, similar to saliency maps, was developed to identify important substructures that influence the biological effect similarity of chemical structure pairs. First, the derivative of deepSIBA's output w.r.t the input matrices that contain the atom features of each compound, in the input pair, is calculated using Tensorflow $\left(\left[\frac{\partial F}{\partial X_{\text{atoms}}}\right]\right)$. Subsequently, for each compound atom importance is scored, using a directional derivative approach. Thus, similar to vector calculus, the directional derivative of a scalar $f(X)$, with X being a matrix, in the direction of a matrix Y is

$$\nabla_Y f(X) = \text{tr}\left(\frac{\partial f}{\partial X} \times Y\right),$$

where, $\frac{\partial f}{\partial X}$ is the gradient matrix, or in our case $\frac{\partial F}{\partial X_{\text{atoms}}}$, while Y can be considered a matrix with zeros everywhere, except the row containing the specific atom's feature. Thus, an importance score for each atom of a compound can be calculated as

$$S_a = \text{tr}\left(\frac{\partial F}{\partial X_{\text{atoms}}} \times Y_a\right),$$

where the only non-zero part of Y_a is the one-hot encoded feature vector of atom a. For each atom the importance score S_a was transformed to a count score C_a, based on how many times each atom was in the top 20% most important atoms for each model in a deepSIBA ensemble. When scoring atom importance during the pathway inference approach, a similar score was calculated based on the times an atom was present in the top 20% for each target-reference pair. Finally, due to the GCNN core module of deepSIBA, important substructures are formed by important atoms that are neighbors in the compound's molecular graph. Atom importance is visualized using the RDKit library.[40]

3 Results and discussion

3.1 Biological factors influence the model's learning task

The presented model is tasked to predict the biological effect distance between compounds, using their molecular graphs as input. Considering that this distance is calculated from experimental GEx data following compound treatment, there are specific biological factors that can influence the learning task. The CMap dataset contains over 110 K transcriptomic signatures from over 20 K compounds assayed across 70 cell lines. By carefully analyzing these signatures and their pairwise distances, we were able to pinpoint the most influential factors and identify their effect on the model's target value.

The variation in quality of GEx data is reflected on the calculated distance value. The quality of gene expression data, from which transcriptomic signatures in the Connectivity map were derived, varies across compound perturbations. In our case, this variation in data quality is especially important. On this front, a categorical quality score, ranging from Q1 to Q8, was assigned to each signature, with a score of Q1 representing the highest quality (see ESI† 1.1). In order to assess the effect of signature quality, distributions of distances between duplicate transcriptomic signatures (same compound, cell line, dose, time) for different quality scores were examined and are presented in Fig. 3A. As expected, Q1 duplicate signatures are very similar and their distances are centered near a small value. However, this is not the case for Q2 duplicate signatures, where differences in differentially expressed genes are prominent even when all the perturbation parameters are kept constant. It is clear that signature quality significantly affects the distribution of the model's target variable.

Distances between transcriptomic signatures vary across cell lines. Compound response, in terms of DEx genes, is highly dependent on the cellular model. Due to different genetic backgrounds and gene expression patterns the same compound perturbation will have different transcriptomic signatures across cell lines.[41] This dependence, directly affects the distance between compounds' transcriptomic signatures for different cell lines. The relationship between gene-level distances of compound pairs present in both the MCF7 and VCAP cell lines, with Q1 signatures, is shown in Fig. 3B. In general, Q1 transcriptomic distances of the same compound pair in the 2 examined cell lines are moderately correlated (Pearson's r = 0.469). However, there is a significant number of compound pairs which have similar transcriptomic signatures in one cell line but not in the

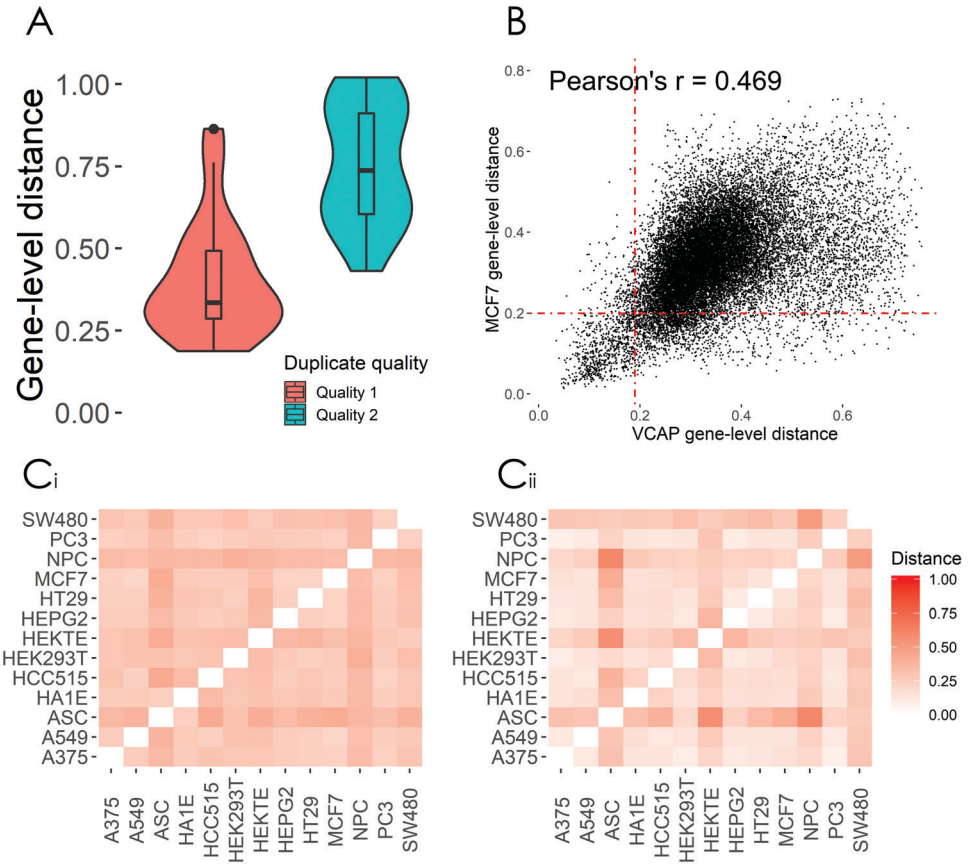

Fig. 3 Influence of biological factors on the learning task. (A) Evaluation of data quality based on the gene-level distance between duplicate compound perturbations (same compounds) for the MCF7 cell line. (B) Scatterplot of distances between transcriptomic signatures (Quality 1) of compound pairs present in both the MCF7 and VCAP cell lines. The red lines, at 0.2 for MCF7 and 0.19 for VCAP indicate the mean + standard deviation of the distribution of distances between Q1 duplicate signatures for each respective cell line; (C_i & C_{ii}) Heatmaps of gene and BP-level distances between cell lines for the knockdown of the MYC gene.

other (lower right and upper right quadrants of Fig. 3B). Such cases are even more prominent for compound pairs with Q2 signatures (see ESI† 1.2). Thus, the cell line effect poses a problem for the proposed learning task by providing a one-to-many mapping between input (pair of chemical structures) and output (distance between signatures).

Compounds' biological effects are better represented on a functional level. A distance function that operates directly on transcriptomic signatures does not account for smaller differences in the DEx of genes that belong to the same biological pathway. Thus, the similar effect between perturbations, in terms of enriched BPs, might not be clearly reflected on their gene-level distance. On this front, a comparison of BP and gene-level distances between cell lines for the knockdown of the MYC gene (Q1 signatures) with shRNA is presented in Fig. 3C. MYC is an oncogene that plays a key role in cell cycle, transformation and proliferation and was selected because its knockdown is expected to cause similar response across cancer cell lines. The smaller overall distance between cell lines in Fig. $3C_{ii}$ indicates that the expected similar effect of MYC knockdown is better highlighted on a functional level between enriched biological processes rather than between transcriptomic signatures (Fig. $3C_i$). Furthermore, we evaluated which distance metric, either between BPs or DEx genes, can better highlight the expected similar biological effect of structurally similar compounds.[42] In the CMap dataset, we identified pairs of similar chemical structure using the traditional Tanimoto coefficient between ECFP4 fingerprints and then calculated what percentage of those cause similar biological response at the BP and gene-level (Table 1). As it can be seen in Table 1, across all structural distance thresholds the percentage of structurally similar compounds with similar biological effect is significantly higher when distance is calculated between signatures of enriched BPs. A detailed comparison between structural and biological effect distances for all examined cell lines is presented in ESI† 1.4.

Through the careful analysis of the processed data sets, we showed that raw data quality greatly affects the distribution of distance values and that lower quality transcriptomic signatures of the same compound, with the same perturbation parameters (duplicates), often exhibit large differences in terms of DEx genes (Fig. 3A). Based on these findings, we chose to develop deepSIBA using only compounds with available Q1 transcriptomic signatures. Furthermore, we showed that the transcriptomic distance of a compound pair can vary depending

Table 1 Percentage of structurally similar compounds that cause similar biological effect, either at the gene or BP-level, in the MCF7 cell line

Structural distance threshold	Pairs with similar chemical structure	Pairs affecting similar BPs[a] (%)	Pairs affecting similar genes[b] (%)
0.10	91	76.9	68.1
0.15	114	75.4	65.7
0.20	200	74.0	61.0
0.25	316	69.9	57.6
0.30	494	65.3	51.0

[a] BP distance threshold to consider compounds similar = 0.2. [b] Gene distance threshold to consider compounds similar = 0.19.

on the choice of cellular model (Fig. 3B). One common approach to address this issue is to aggregate either transcriptomic signatures or distance values across cell lines. While aggregating enables the training of a general model on all available compound pairs, it can often produce misleading results and cause information loss. Thus, we decided to make our approach cell line specific and develop our models for cell lines that have the highest number of Q1 transcriptomic signatures following compound treatment. Finally, we highlighted that a distance function operating on enriched BPs, rather than genes, can better capture the expected biological effect similarities of perturbations with similar structure or biological nature (Table 1 and Fig. 3C). We reason that this is the case due to the BP enrichment analysis that precedes the distance calculation, which can capture smaller changes in the expression of genes that interact with each other to form a biological process. By analyzing the relationship between the aforementioned experimental factors and our target variable, we were able to make data-driven decisions to propose a learning task that minimizes their effect. In the following sections we evaluate the ability of deepSIBA to learn the proposed task and test whether our approach can identify dissimilar structures that affect similar BPs in a meaningful way.

3.2 Performance evaluation

Model performance was evaluated on pairs of reference and test compounds. Test compounds were removed from the training sets and thus represent new chemical structures without available experimental GEx data. Additionally, the effect of the structural similarity between input compounds on performance, along with the utility of the model's estimate for uncertainty, were investigated. Finally, we evaluated the performance of our approach on test chemical structures that are very different from the ones used in training.

Performance evaluation in cell line specific test sets. In each cell line specific test set, the performance of deepSIBA was compared to the performance of ReSimNet and TwoStepRLS. ReSimNet is a recently proposed deep Siamese MLP model, while TwoStepRLS is a regularized kernel-based regression method. Both methods are suitable for distance/similarity learning for pairwise (dyadic) data and were implemented using compounds' ECFP4 fingerprints as input (see ESI† 3.1 and 3.2 for details). As shown in Table 2, across all cell lines, deepSIBA achieved the lowest overall MSE and in the 1% of test samples with the lowest predicted values. The ReSimNet models for the A375 and MCF7 cell lines achieved the highest Pearson's r, while deepSIBA and TwoStepRLS had the highest Pearson's r, for the PC3 and VCAP cell lines respectively. In terms of precision, the deepSIBA models heavily outperformed the other methods across all cell lines. In order to calculate precision, an appropriate distance threshold of 0.2 was used for all approaches (see Section 2.5 for details) While ReSimNet and TwoStepRLS exhibited low precision, they predicted that many more compound pairs will have similar biological effect. When examining the lowest 1% of predicted distances, their precision improves and in the MCF7 cell line TwoStepRLS' precision surpasses deepSIBA's. Additional 5-fold cross validation results for each cell line are presented in ESI† 5.2.

Transferring knowledge to other cellular models. Initially deepSIBA was trained and evaluated in the four cell lines that have the highest number of Q1 transcriptomic signatures following compound treatment. In order to expand the biological coverage of deepSIBA we utilized transfer learning to train our models on six additional cell lines which have the next highest number of Q1 signatures. On this front, we pretrained a deepSIBA model on the entirety of the A375 cell line dataset and then applied it on additional cell lines by resuming

Table 2 Cell line specific test set performance

Cell line	Model	MSE	MSE @1%	Pearson's r	Precision (%)	Precision @1% (%)	Predicted similar pairs
A375	DeepSIBA	**0.008**	**0.006**	0.59	**98.22**	**98.22**	169
	ReSimNet	0.012	0.022	**0.60**	32.23	56.80	18 243
	TwoStepRLS	0.010	0.008	0.51	44.61	78.68	4024
PC3	DeepSIBA	**0.011**	**0.007**	**0.53**	**89.29**	**89.29**	28
	ReSimNet	0.017	0.032	0.49	25.02	46.89	14 195
	TwoStepRLS	0.013	0.041	0.44	29.98	38.96	1758
VCAP	DeepSIBA	**0.033**	**0.026**	0.41	**71.63**	**71.63**	141
	ReSimNet	0.039	0.105	0.38	32.69	52.97	9245
	TwoStepRLS	0.034	0.049	**0.43**	32.34	31.12	3120
MCF7	DeepSIBA	**0.012**	**0.007**	0.56	**61.03**	61.03	195
	ReSimNet	0.015	0.029	**0.59**	26.93	51.20	13 420
	TwoStepRLS	0.015	0.010	0.47	33.55	**70.14**	4322

training for 6 epochs. The performance of the transfer learning approach on each cell line specific test set is presented in Table 3. Across all additional cell lines deepSIBA was able to achieve similar performance to that of the A375, PC3, MCF7 and VCAP cell lines.

Performance as a function of the structural distance between input compounds. As shown previously, similar chemical structures have similar signatures of enriched BPs. However, there are many cases of structurally dissimilar compounds that cause similar biological response. It is therefore important to evaluate the ability of deepSIBA to identify such cases, by calculating its performance for test pairs of varying structural distance. On this front, each cell line specific test set was split into parts based on the structural distance between compounds and in each part MSE and precision were calculated (Fig. 4A and B). As a measure of structural distance/similarity, the traditional Tanimoto coefficient between ECFP4 fingerprints was utilized. The PC3, A375 and VCAP deepSIBA models maintain a high precision across all different structural distance ranges (Fig. 4B). The exception is the MCF7 model, for which precision slightly decreases for structural distance higher than 0.7. Regarding MSE, only the VCAP model exhibits a slightly higher MSE as structural distance increases (Fig. 4A). As a whole, the models' performance seems unaffected by the distance between the ECFP4 fingerprints of the input pairs.

Performance as a function of predictive uncertainty. It has been shown that quantifying predictive uncertainty can lead to more accurate results in virtual screening applications.[30] In this context, the relationship between the predictive uncertainty estimate and performance was investigated. In DeepSIBA we estimate predictive uncertainty as the coefficient of variation (CV) of the mixture of each model's Gaussian in the ensemble. MSE and precision were calculated for specific samples in the test set, which have CV lower than an increasing threshold and are presented in Fig. 4C and D. As the CV threshold increases and more samples with higher CV are included in the evaluation, the MSE of the models increases as well and eventually becomes the MSE of the entire test set (Fig. 4C). On the other hand, due to the low number of false positives, for all the models, precision seems unaffected by the CV threshold. Only the MCF7 model, which has the lowest overall precision, exhibits a higher precision for samples with lower CV (Fig. 4D). Overall, the results indicate that point predictions with lower uncertainty are closer to the true value, or that when the model is certain, it's usually not wrong.

Generalization on different chemical structures. End-to-end deep learning models for drug discovery have trouble generalizing

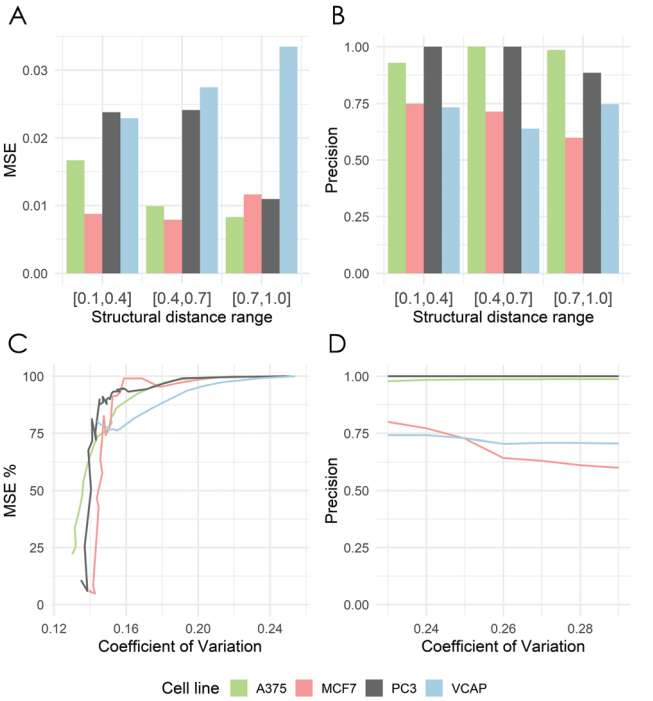

Fig. 4 Performance as a function of structural distance and predictive uncertainty; (A) MSE for different ranges of structural distance between compound pairs. (B) Precision for different ranges of structural distance between compound pairs. (C) Percentage of total test MSE, calculated in samples with increasing CV. (D) Precision calculated in test samples with increasing CV.

on new compounds that are structurally very different from the ones used to train them. In order to evaluate the ability of our approach to generalize on different chemical structures, the performance of the A375 model was evaluated on 2 extra test sets and is presented in Table 4. These test sets were created by restricting the maximum allowed structural similarity between selected test compounds and all remaining training compounds and thus represent test scenarios of increasing difficulty (Fig. 5A). As the minimum distance between test and training compounds increases, the performance of the model becomes worse. However, the performance decrease in terms of MSE and Pearson's r is smaller than the decrease in precision. In this case, the distance threshold to calculate precision was set to 0.22, because in the hardest test set (#3) there were no samples with predicted value lower than 0.2. Thus, even though the model's performance is comparable across test sets in terms of regression metrics, its ability to identify compounds with similar biological effect is hindered. In this case, it is important to

Table 3 Test set performance of the transfer learning approach

Cell-line	MSE	MSE @1%	Pearson's r	Precision (%)
HT29	0.010	0.013	0.60	84.88
A549	0.013	0.012	0.62	83.00
HA1E	0.015	0.009	0.58	100
HEPG2	0.013	0.014	0.61	99.10
HCC515	0.014	0.010	0.52	97.92
NPC	0.006	0.005	0.67	73.64

Table 4 Generalization performance on different chemical structures for A375

Test set	Max similarity to training set	MSE	Pearson's r	Precision (%)	Predicted similar pairs
#1	[0–0.85]	0.0083	0.59	97.26	876
#2	[0.35–0.65]	0.0092	0.52	76.48	330
#3	[0–0.3]	0.0107	0.44	50.37	135

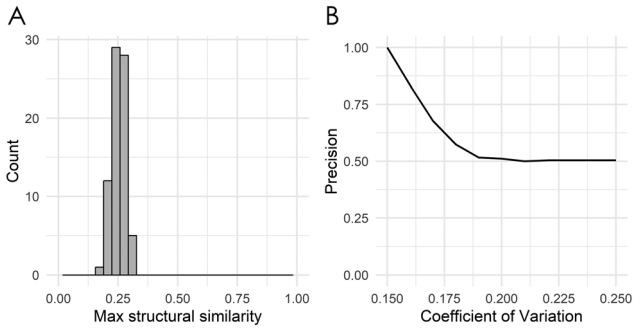

Fig. 5 Precision and uncertainty estimation for test set number 3. (A) Histogram of maximum structural similarity between test and training compounds for test set number 3; structural similarity is calculated between compounds' ECFP4 fingerprints. (B) Precision calculated in test samples with CV lower than an increasing threshold.

estimate predictive uncertainty and evaluate its utility, by focusing on predictions with smaller CV (Fig. 5B). In the third test set, which only contains compounds with maximum similarity to the training compounds less than 0.3, the model's precision is significantly higher for test predictions with low CV. More specifically, in test samples with CV lower than 0.16, the model's precision is upwards of 80%.

Across all examined cell lines, deepSIBA was able to identify chemical structures that affect similar BPs, outperforming, especially in terms of precision, the distance learning methods that utilize compounds' ECFP4 fingerprints as input (Table 2). Even though the learning task is regression, we reason that precision is a crucial metric, considering the potential screening applications of deepSIBA in order to identify compounds that exhibit similar biological effect to a query. In this scenario, high precision, rather than a large number of identified hits, is required to correctly prioritize compounds for downstream experimental validation. We chose not to compare our approach with traditional machine learning methods, *e.g.* Random Forests and SVMs, because we argue that these are not optimal for a distance/similarity learning task. Furthermore, deepSIBA was able to maintain its high performance regardless of the structural similarity between input compounds and identify cases of structurally dissimilar compounds that affect similar BPs (Fig. 4A and B). Thus, the employed GCNN architecture shows promise towards this highly interdisciplinary task. However, there were some cases of compounds affecting similar BPs that were missed by the model. These cases, in combination with the decrease in performance as the minimum structural distance between test and training compounds increases highlight key limitations in our approach (Table 4). On this front, limited coverage of the chemical space by compounds with available GEx data is a major issue that limits our ability to model in its entirety the complex function that translates changes in chemical structure to BP alterations. Even though each training set for each cell line contains on average around 320 K samples, these are comprised from the pairing of around 800 compounds. The limitations that arise from this low coverage of the chemical space can't be solved by changes in deep learning architecture and require more training compounds and/or extra input information. On this front, we applied a data augmentation technique, where each training set was augmented with randomly sampled pairs between Q1 and Q2 compound signatures (see ESI† 4.2). However, due to conflicting evidence between Q1 and Q2 transcriptomic signatures the performance of the models varied significantly across cell lines (see ESI† 5.3). A rather efficient workaround that we utilized in our approach is to quantify predictive uncertainty using deep ensembles. We showed that the model's performance, even when tested on compounds that are structurally different from the ones used in training, is higher for samples with lower uncertainty (Fig. 5). Thus, the model's estimate of predictive uncertainty can be used to provide more reliable and accurate results. For instance, if an application imposes a constraint on the maximum allowed error, the appropriate uncertainty threshold can be identified and only point predictions with uncertainty lower than this threshold can be considered. Finally, we showed that transfer learning is a suitable approach to expand the biological coverage of deepSIBA to additional cellular models with fewer available data points (Table 3). For example, in the NPC cell line, which has approximately 50% fewer compound signatures than A375, deepSIBA was still able to achieve reasonable performance.

3.3 Signaling pathway inference for target structure

The predictions of deepSIBA can be used to infer a signaling pathway signature, in terms of the most upregulated and downregulated pathways, for a target chemical structure without available GEx data. The inference is performed following a KNN-like approach, in which reference compounds with the smallest distance to the target, as predicted by the model, are selected as its neighbors and their pathway signatures are retrieved. Then, pathways that frequently belong in the 10 most upregulated or downregulated pathways of the neighbors are inferred as the target's signature. The performance of the approach was evaluated on the test compounds of the MCF7 model and then, as a use case, it was tasked to infer the signaling pathways affected by FDA approved anticancer drugs, for which no GEx data are available in our dataset. Additionally, the chemical substructures that mostly influence the inferred pathways were identified and visualized using a graph gradient-based approach.

Evaluation of the pathway inference approach in the test set of MCF7. For the test set of the MCF7 cell line, the average performance of the inference approach is presented in Table 5. On average 5 pathways per test compound were inferred to belong in its 10 most downregulated pathways with a precision of 73.3%. Regarding upregulation, an average of 2.5 pathways per compound with a precision of 69.7% were inferred. We have to note that the statistical significance of the inferred pathways is ensured by comparing the neighbor selection process using the trained model to a random selection.

Use case: signaling pathway inference of FDA approved anticancer drugs. Out of the 59 FDA-approved cytotoxic drugs presented by Sun *et al.*, 18 were present or had a structural analogue in the MCF7 training set (Tanimoto ECFP4 similarity

Table 5 Pathway inference results for the test compounds of MCF7

	Number of inferred pathways	Precision (%)
Downregulated	5	73.3
Upregulated	2.5	69.7

>0.85).[43] In order to simulate a realistic application for the signaling pathway inference, these 18 drugs were excluded from the use-case. From the remaining 39 drugs, only 3 had more than 5 neighbors each in the training set, as predicted by the model and the inferred pathways are presented in Table 6. Fludarabine and Clofarabine are direct nucleic acid synthesis inhibitors, while Pralatrexate is an indirect inhibitor of nucleotide synthesis through inhibition of the folate cycle.[44] In our use case, the inferred downregulated signaling pathways include cell cycle, purine and pyrimidine metabolism, RNA transport and spliceosome, which are closely related to the 2drugs' mechanism of action. Furthermore, because of the MCF7 cell line, pathways such as oocyte meiosis and progesterone-mediated oocyte maturation, that have been associated with the pathogenesis of breast cancer, were inferred as downregulated.[45] Regarding upregulation, pathways such as NF-kappa B signaling, natural killer cell mediated cytotoxicity, leukocyte transendothelial migration and TNF signaling, that are closely related to inflammation and apoptosis, were inferred.

Important substructure identification for the drugs in the use case. The method described in Section 2.7 was used to highlight important substructures that deepSIBA pays attention to when inferring the pathway signature of each anticancer compound presented in the use case (Table 6 and Fig. 6). In Fig. 6, red colored atoms represent atoms for which the model exhibits large directional derivatives across all pairs of target and neighbor compounds. Such atoms that are closely connected in the target compound's molecular graph are identified as influential to the inferred pathway signature. As shown in Fig. 6, for Fludarabine and Clofarabine, deepSIBA highlights the 2-fluoroadenine and 2-chloroadenine substructures as important respectively, while the model mostly focuses on the Pteridine structure when inferring the pathways affected by Pralatrexate.

In the presented use case, we demonstrated that by utilizing the training compounds as reference, the inferred signaling pathway signatures for each of the anticancer drugs were found to be closely connected to their respective MoA (Table 6). Thus, our inference method has the potential to provide an early estimate regarding the pathways affected by a compound, using only its chemical structure as input. Additionally, we showed that for each compound the highlighted substructures are also directly related to their respective MoA (Fig. 6). This fact not only increases the interpretability of the model's predictions, which is a crucial topic of DL methods for drug discovery, but also shows that a GCNN model trained end-to-end on molecular graphs is able to learn meaningful structural representations that are related to compounds' biological effects.[46–48] To the best

Table 6 Pathway inference results for FDA approved anticancer drugs

Drug	Mechanism of action	Inferred downregulated KEGG signaling pathways	Inferred upregulated KEGG signaling pathways
Fludarabine	Nucleic acid synthesis inhibitor	***Purine metabolism, pyrimidine metabolism***, **RNA transport**, **spliceosome**, **cell cycle**, **oocyte meiosis**, **progesterone-mediated oocyte maturation**, MicroRNAs in cancer	Leukocyte transendothelial migration, oxytocin signaling pathway, Alzheimer's disease, pertussis, rheumatoid arthritis
Clofarabine	Nucleic acid synthesis inhibitor	**RNA transport**, **spliceosome**, **cell cycle**, ubiquitin mediated proteolysis, **progesterone-mediated oocyte maturation**, MicroRNAs in cancer	**Natural killer cell mediated cytotoxicity**, leukocyte transendothelial migration, oxytocin signaling pathway, pertussis, rheumatoid arthritis
Pralatrexate	Inhibits dihydrofolate reductase (DHFR) and thymidylate synthase	**Purine metabolism**, **pyrimidine metabolism**, metabolic pathways, **RNA transport**, **spliceosome**	**NF-kappa B signaling pathway**, **natural killer cell mediated cytotoxicity**, **TNF signaling pathway**, leukocyte transendothelial migration

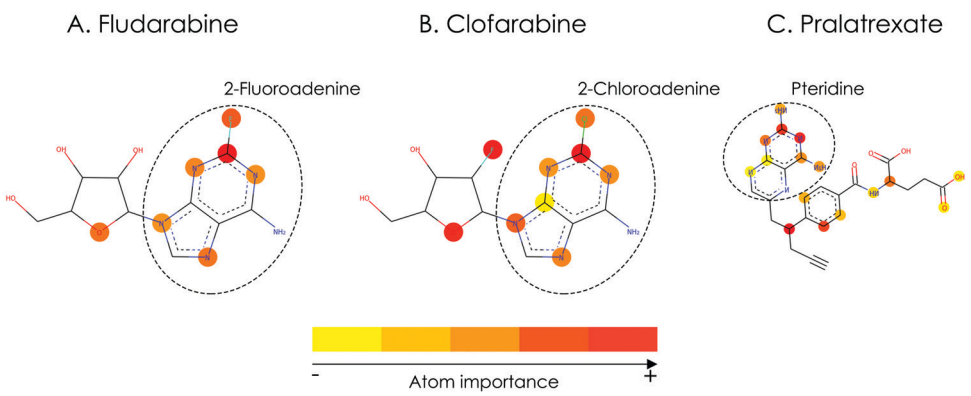

Fig. 6 Important atoms related to the inferred biological footprint of the compounds of the use case, as identified by the deep learning model (the red color signifies the most important atoms).

of our knowledge, this is the first time a DL model was used to identify important substructures and infer the signaling pathway signature of a target compound without available experimental GEx data. A possible limitation of our approach might be its resolution capabilities in specific use-cases of compounds with similar chemical structure but different MoA. Although the comparison of the fludarabine and clofarabine use-cases suggests that our approach might be able to identify small structural differences between drugs with similar MoA (Fig. 6), we haven't systematically compared use-cases of structurally similar compounds that affect different BPs. From the analysis of the CMap dataset we have showed that compounds with high structural similarity tend to have similar biological effect (see ESI† Fig. S6). This lack of data regarding compounds that are derivatives but affect different BPs limits our ability to systematically perform the aforementioned comparison and pinpoint the maximum resolution of our approach. Furthermore, due to the nature of the inference method, limiting factors may also arise from the lack of diversity in affected BPs by the training compounds. This lack of diversity can influence the signaling pathway inference for an unknown target structure, when its true biological footprint is not represented in the reference compounds. In such cases, the inference of incorrect signatures can be avoided by focusing on target compounds with at least k reference neighbors (here $k = 5$) and only infer statistically significant pathways, using our method's calculated p-value.

4. Conclusion

In this paper, we developed a deep learning framework to match the chemical structure of compound perturbations to their biological effect on specific cellular models. We showed, that the careful formulation of the learning problem and the flexibility of the Siamese GCNN architecture enabled our models to achieve high performance across all test scenarios. Additionally, we highlighted the utility of the uncertainty estimate, provided by deep ensembles, in test cases where the unknown chemical structures are very different from the structures used to train the models. Finally, we presented a novel inference pipeline, which can infer a signaling pathway signature for a target compound and subsequently identify which substructures mostly influenced the prediction. The novelty, performance and interpretability of our methods paves the way for further investigation in order to expand their coverage and utility.

Possible efforts for further investigation can be concentrated on the input representation, the biological response distance and the model's uncertainty estimate. Regarding the input, one interesting idea is to include binding information in order to capture the potential protein target of the input molecules. This extra information can be passed to the model either in the form of latent space embeddings from a trained binding affinity prediction model or in the form of predictions against a panel of protein kinases.[49] Regarding the biological distance between compound perturbations, this can be augmented by calculating the compound's effect on different levels of biological hierarchy, *i.e.* GEx, signaling pathways, transcription factors and signaling networks.[50,51] Afterwards, these distances could be combined or separate models could be trained in order to better capture the similar effect of compounds. Additionally, instead of using a distance metric between all affected BPs, specific biological processes could be selected and application specific models could be developed to identify compounds that affect these biological processes. Regarding the model's uncertainty estimate, an interesting avenue for investigation is to take into account the transcriptomic signatures of replicates from the CMap dataset and calculate distributions of pairwise distances between compounds. Then, models could be trained on these distributions to better capture the variation of the experimental ground truth. Finally, collecting more data regarding derivative compounds with different MoA is an interesting avenue for further investigation in order to identify the resolution capabilities of the substructure importance approach.

The highly interdisciplinary framework of deepSIBA combines aspects from both the CADD and 'omics domains in order to incorporate the structural and systematic effects of small molecule perturbations, which are closely related to their efficacy and toxicity profiles. We believe that our methods have the potential to augment *in silico* drug discovery, either by exploring on a massive scale the biological effect of compounds/libraries without available GEx data, or by suggesting new chemical structures with desired biological effect.

Data and code availability

All analyzed data that were used to train our models and produce all tables and figures are available at https://github.com/BioSysLab/deepSIBA. Furthermore, the R source code to analyze the CMap dataset and create the training, validation and test sets is available at https://github.com/BioSysLab/deepSIBA/preprocessing. Finally, the Keras/TensorFlow implementation of our deep learning models, alongside trained ensemble models for each cell line are available at https://github.com/BioSysLab/deepSIBA/learning.

Author contributions

C. F. and N. M. conceived the idea and trained the models. C. F., N. M. and A. S. preprocessed the data. L. G. A. supervised the project. All authors analyzed the results and wrote the manuscript together.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We would like to thank T. Sakellaropoulos and M. Neidlin for their constructive feedback on the manuscript and N. Emmanouilidi for her help in designing the figures. Finally, we would like to

