Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Genome analysis

Deep GONet: Self-explainable deep neural network using Gene Ontology for phenotype prediction from gene expression data

Victoria Bourgeais [1,*], Farida Zehraoui [1], Mohamed Ben Hamdoune [1] and Blaise Hanczar [1,*]

[1] IBISC, Université Paris-Saclay (Univ. Évry), Évry, 91034, France.

* To whom correspondence should be addressed.

Abstract

Motivation: With the rapid advancement of genomic sequencing techniques, massive production of gene expression data is becoming possible, which prompts the development of precision medicine. Deep learning is a promising approach for phenotype prediction (clinical diagnosis, prognosis, and drug response) based on gene expression. Existing deep learning models are usually considered as black-boxes that provide accurate predictions but are not interpretable. However, accuracy and interpretation are both essential for precision medicine. Hence, making deep learning models interpretable for medical applications is the main focus of this paper.

Results: In this paper, we propose an innovative Deep GONet model, integrating the Gene Ontology into the hierarchical architecture, which is self-explainable. This model is based on a fully-connected architecture constrained by the Gene Ontology annotations, such that some informative genes, which have not been defined biologically yet, can be used. The experiments on cancer diagnosis dataset demonstrate that Deep GONet is both easily interpretable and highly performant to predict cancer regardless of the type of cancer. We succeed to explain the most important neurons involved in cancer predictions by their associated biological functions, making the model understandable for biologists and physicians.

Availability: Deep GONet is implemented in Python using the Tensorflow framework. The code is available on https://forge.ibisc.univ-evry.fr/vbourgeais/DeepGONet.git.

Contact: blaise.hanczar@ibisc.univ-evry.fr, victoria.bourgeais@univ-evry.fr

Keywords: Gene expression, Phenotype prediction, Model interpretation, Deep learning, Gene ontology.

1 Introduction

With the rapid advances of data acquisition technologies, collecting large amounts of different-type data (images, ECG, genomics…) becomes simpler in the medical field. It inspires a new form of this field, i.e., precision medicine, which takes advantage of these available data to improve profoundly diagnosis, prognosis, or therapeutic decision. Precision medicine has access to detect in advance a disease, such as cancers, anticipate the progression of the disease, and adapt the therapy according to the characteristics of patients. Among these data, genomic data and especially gene expression play a key role in the development of precision medicine. Gene expression profile is known to be an indicator of the cellular state and allows the study of complex diseases.

For many years, machine learning has been used on transcriptomic data to construct classifiers predicting phenotypes from gene expression profiles (Kourou *et al.*, 2015). In the last decade, deep learning has become the source of the most impressive improvements in machine learning (Goodfellow *et al.*, 2016). It shows its superiority in many domains such as image analysis or natural language processing. For the last two years, deep learning begins to be applied to classification based on gene expression problems. Unlike images or texts, gene expression data have no structure. The architectures used in literature are therefore mainly the autoencoders and multilayer perceptrons (MLP) (Zou *et al.*, 2019). For instance, *Stacked Denoising Autoencoders* (Fakoor *et al.*, 2013; Danaee *et al.*, 2017) are exploited to extract a lower dimension from the data, then a classifier (such as linear model, support vector machine (SVM) or MLP) is applied to accomplish classification. Multilayer perceptrons are used in (Guo *et al.*, 2017; Hanczar *et al.*, 2018) to predict directly diseases

without dimension reduction. Today, despite promising first results, deep learning has not made a breakthrough in gene expression classification yet because of the small size of the available training sets. Deep learning is especially good with large training sets. In the next years, with the increasing production of transcriptomic data, it is highly likely that deep learning will play a major role in these problems.

One of the main concerns of deep learning in medical applications is its lack of interpretability. Indeed, the neural networks are black-box models, which means that the model cannot provide an explanation to its decision. The interpretation of machine learning algorithms became a critical topic in the last decade and especially in the case of medical application for three reasons. First, both the physician and his patient must understand why the model predicts a given phenotype (diagnosis, prognosis, treatment). Particularly, it can influence later decisions. Second, it is important to ensure that the model bases its predictions on a reliable representation of the data and does not focus on irrelevant artefacts. This will highly impact the trust of the physicians toward the predictions regardless of the performances of the model. Finally, the model with high accurate predictions may have identified interesting patterns that biologists would like to investigate. We can distinguish two main approaches for interpreting the black-boxes: the post-hoc methods and the self-explainable models (Adadi and Berrada, 2018).

In a post-hoc method, the black-box model is first learned and then an interpretation method is used to explain the predictions. Several post-hoc methods with different purposes are proposed in the literature. Among them, proxy methods, which approximate a black-box model by an interpretable model, can help interpret the general behavior of the model. For example, Ribeiro *et al.* propose a linear proxy method, *Local Interpretable Model-Agnostic Explanations* (LIME), to approximate any black-box model (Ribeiro *et al.*, 2016). Interpretation methods specific to deep learning have been recently proposed (Ancona *et al.*, 2019). The model prediction is explained by backpropagating the signal from the output to the input. This type of method enables the identification of the most relevant features and neurons involved in decision making. Several gradient-based methods are proposed in the literature including *Layerwise Relevance Propagation* (LRP) (Bach *et al.*, 2015; Montavon *et al.*, 2017), *Integrated Gradients* (Sundararajan *et al.*, 2017) and *DeepLift* (Shrikumar *et al.*, 2017). In (Ancona *et al.*, 2019), authors present a unified framework for interpreting predictions by analyzing several gradient interpretation methods from theoretical and practical perspectives. They show that these methods are strongly related and equivalent under certain conditions. Lundberg and Lee show that among gradient-based methods, DeepLift and LRP are better aligned with human intuition as measured by studies since they satisfy some desirable properties (Lundberg and Lee, 2017).

The self-explainable models are inherently interpretable models. The interpretability is a characteristic of this family of models. The most famous methods of this family are the decision trees, rules systems, sparse linear models. Few works on self-explainability have been proposed for deep learning. Melis and Jaakkola introduces a built-in interpretable model, *Self-Explainable Neural Network*, that behaves locally as a linear model (Melis and Jaakkola, 2018). A general opinion is that the black-boxes are more accurate than the self-explainable models. The capacity of interpretability is often viewed as a constraint of the model that decreases its performance. There would be a tradeoff between the performances and interpretability. However, recently a part of the machine learning community claims that performance and interpretability are not exclusive. Rudin explains why black-box models should be avoided for crucial decisions, like in medical applications, even with the use of post-hoc interpretation (Rudin, 2019). They promote the development of high-accurate self-explainable models.

All of these methods, post-hoc and self-explainable, consider that the interpretation of a model is the identification of the inputs and the part of the model, in case of deep learning, the set of neurons which support the predictions. It is a mathematical interpretation of the predictions which generally does not provide an understandable explanation. The explanation must be completed with knowledge of the domain. In the case of phenotype prediction from gene expression, we have to explain which biological notion is represented in the model and which one is used to compute the predictions. Few works have been published on the construction of self-explainable models based on gene expression data. Prior knowledge comes from the ontologies such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Snel *et al.*, 2000) or Gene Ontology (GO) (Consortium, 2004). Among them, the closest literature to our work (Peng *et al.*, 2019) incorporates GO into a neural network, defined as *Gene Ontology Neural Network* (GONN). They replace one hidden layer by one level of the GO ontologies and connect the input features by partial connections according to the connections with the ontologies. In this way, some input features cannot be included if there are not connected to the ontologies. Similarly, *Pathway-Associated Sparse Deep Neural Network* (PASNet) (Hao *et al.*, 2018) and *Gene Regulatory network-based Regularized Artificial Neural Network* (GRRANN) (Kang *et al.*, 2017) integrate respectively biological pathways and regulators from Protein-protein/Protein–genetic interactions in the first layer. These architectures contain at most two hidden layers. For example, PASNet tries to capture nonlinear interactions between pathways in a second hidden layer. However, deep neural networks allow deeper representations of hierarchical relations between gene expression and biological objects. By using prior knowledge, these works first attempt to boost the model performances on their target tasks. Yet, they do not clearly show whether the neurons correspond to the associated biological concept or not. Integrating knowledge may not be so beneficial for learning.

In this paper, we propose a self-explainable neural network, called *Deep GONet*, based on gene expression data. This model is constrained by prior biological knowledge from GO, which is widely used in the bioinformatics community. The architecture represents different levels of the ontology preserving the hierarchical relationships between the GO terms using sparse regularization. Our objective is to build an accurate and relevant interpretable model for cancer detection. Each neuron is associated with a GO function and the links between these functions are represented by the network connections. A prediction of the network can, therefore, be directly explained by the set of biological functions.

The paper is organized as follows. In Section 2, we describe the proposed novel model Deep GONet for biological interpretation. In Section 3, the model is evaluated using gene expression data and compared with other approaches. Section 4 provides how to obtain the explanations of a prediction and their biological significations. The conclusions to this paper and some future research directions are presented in Section 5.

2 Methods and Materials

We propose a new neural network model, Deep GONet that is self-explainable and embeds the biological knowledge contained in GO. Our model is based on a multilayer perceptron constrained by the GO structure. The constraints are introduced into the network using an adaptive regularization term.

2.1 The architecture of Deep GONet

The architecture of our deep neural network represents the structure of GO. GO gathers three ontologies that respectively describe the following categories: biological process (GO-BP), molecular function (GO-MF), and cellular component (GO-CC). We chose to base the architecture of our network on the GO-BP since it provides the most understandable interpretation. However, it is possible to implement the GO-MF or GO-CC in the network architecture with the same method.

GO-BP is structured as a directed acyclic graph containing 11991 nodes distributed over 19 levels as illustrated in the top of Fig. 1. Each node is a GO term representing a biological function. Two GO terms are linked if their biological functions are related, and the majority of these relations are "is a" relations. The GO terms are connected respecting a hierarchical bottom-up orientation. A GO term is assigned to a dedicated level according to its longest path to the root (i.e., GO:0008150). The GO terms in lower levels correspond to more specific functions, like positive regulation of skeletal (GO:0014810 at the 19th level), whereas the nodes in upper levels are more general functions such as the biological process function (GO:0008150). The GO terms are also linked to genes via annotations. A parent GO term (i.e., destination of incoming connections) inherits, therefore, the set of genes from its children (i.e., origins of incoming connections).

Our neural network architecture represents the GO-BP, i.e., each hidden layer l represents a GO level h, each neuron a GO term, and each input a gene. Since the lowest levels of GO contain few very specific terms, and the highest levels are very general, it seems not useful to implement the whole GO in our architecture. We chose to use only level 7 to level 2 (inside the green box in Fig. 1). Level 7 is picked since it is connected to the highest number of genes.

Our model is based on a MLP that consists of an input layer, L hidden layers, and an output layer for phenotype prediction. The input layer is composed of probes. A probe is a short DNA sequence targeting a region of one or several genes. It is the measure used in micro-array data. Each neuron is connected to all neurons of the previous layer and all neurons of the next layer. The output layer returns the probabilities to belong to each class. Each hidden layer corresponds to a level in GO-BP, and its neurons match all the GO terms of the target level. Note that the incorporation of the knowledge must respect the goal of the neural network to construct an abstract representation of the data through its hierarchical architecture. The first hidden layer of a neural network extracts the low-level features from the input layer, its corresponds to level 7 of GO containing more specific biological functions. In the last hidden layers, the gene expression is represented by high-level features representing the most general biological functions of the highest GO levels. The bottom of the Fig. 1 illustrates our method where the levels 7 to 2 of GO are implemented in the architecture our the neural network.

The activation of the i-th neuron of the hidden layer l can be expressed as: $a_i^{(l)} = f\left(\sum_{j=1}^{N_{l-1}} a_j^{l-1} w_{ji}^{(l)} + b_i^{(l)}\right)$, where $w_{ji}^{(l)}$ is the weight of the connection from the j-th neuron of the layer $l-1$ to the i-th neuron of the layer l, $b_i^{(l)}$ is the bias of the i-th neuron of the layer l, N_{l-1} the number of neurons in the layer $l-1$ and f is the rectified linear unit function (ReLU) defined as $f(x) = \max(0, x)$. The output layer contains for binary problem only one neuron with a sigmoid function (as illustrated in Fig. 1) and for multiclass problem one neuron for each class with a softmax function.

In our fully-connected architecture, we identify two types of connections :

- connections corresponding to links in GO-BP (colored in red in Fig. 1), called GO connections;
- connections between two nodes that are not linked in GO-BP (marked by dashed arrows in Fig. 1), called noGO connections.

The probes in the input layer are connected to the neurons of the first hidden layer via a GO connection if it is associated with the corresponding GO term in level 7 of GO-BP, or via a noGO connection otherwise. Note that the neurons of the hidden layer 2 to 6 are not directly connected to the probes. These neurons can be indirectly connected to their probes by the propagation of gene expression through the GO connections of the previous layers. If we want to represent exactly the GO-BP, we can cut all

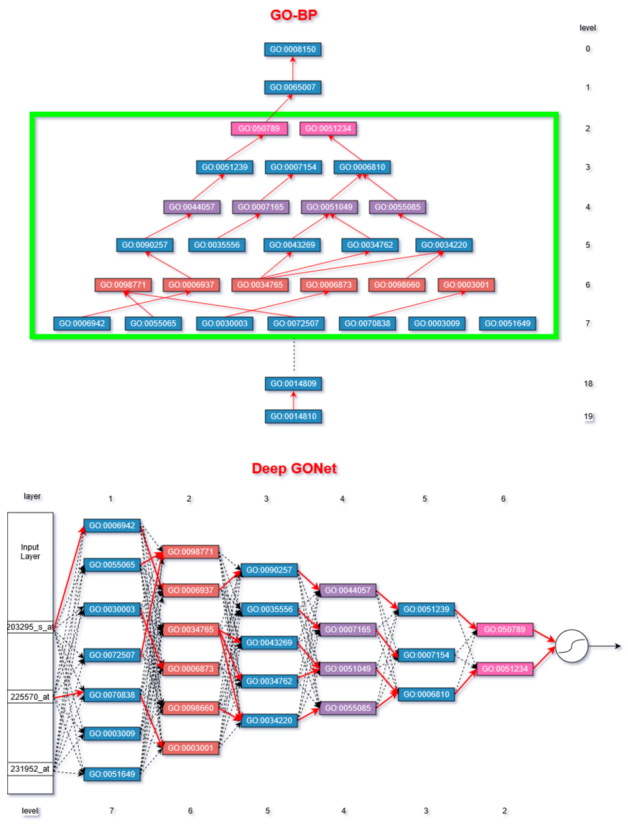

Fig. 1. A subset of GO-BP (top) and the corresponding Deep GONet architecture (down). The green box represents the GO levels implemented in Deep GONet. The red and black dashed arrows represent respectively the GO and noGO connections.

noGO connections and keep only the GO connections in our architecture. However, GO only represents the current knowledge we have on biology. The ontologies change continuously with the outcoming of new scientific discoveries. Some links can be missing or wrong, and many probes are not associated with their right corresponding GO term. 33% of the probes from the microarray HG-U133Plus2 used in our experiments have noGO connections (such as the probe 231952_at in Fig. 1). This means that these probes would not be connected to the neural network if we use only the GO connections. They would not be used to compute the prediction even if they bring relevant information. This situation could impact negatively the accuracy of the neural network. To deal with the errors and the incompleteness of the knowledge represented in GO, we keep all connections in our architecture. However, the noGO connections are penalized to favor the use of GO connections to compute the predictions.

2.2 Regularization of the network

The model is constrained by a customized regularization term, named L_{GO}, to favor the GO connections and penalize the noGO connections. This regularization term is defined as follows:

$$L_{GO} = \sum_{l=1}^{L} \|W^{(l)} \otimes (1 - C^{(l)})\|_2^2 \quad (1)$$

where L is the number of hidden layers of the neural network, $W^{(l)}$ is the weight matrix of the layer l and $\otimes$ is the pointwise product. $C^{(l)}$ is the adjacency matrix that encodes the connections between the GO terms of the layer $l-1$ and l (i.e., the corresponding levels $h-1$ and h in GO-BP). More precisely, if a GO term i at the corresponding level h in GO-BP is

a parent of GO term j from the level $h-1$, then $c_{j,i}^{(l)}=1$ else $c_{j,i}^{(l)}=0$. For the output layer, $C^{(L)}$ is a matrix of ones. The loss of our model is the sum of the common cross-entropy loss and our regularization term :

$$L=\sum_{i=1}^{N}\sum_{k=1}^{K}\left(-y_{i,k}\log\hat{y}_{i,k}\right)+\alpha L_{GO}$$

where N and K are respectively the number of examples in the training set and the number of classes. $y_{i,k}$ is the indicator of the true class, i.e., $y_{i,k}=1$ when the i-th example belongs to the class k, or 0 otherwise. $\hat{y}_{i,k}$ is the probability that the i-th example belongs to the class k computed by the neural network. α is a hyperparameter that weights the regularization term. With $\alpha=0$, the regularization term vanishes our model becoming a classical MLP with no interpretation capacity. With high value of α, the learning algorithm focuses on the regularization term and ignores the cross-entropy. The resulting neural network represents perfectly the GO connections by cutting the noGO connections, but it has no prediction capacity. α is a crucial hyperparameter that controls the trade-off between the interpretability and the performance of our model.

3 Experiments and Results

3.1 Dataset

The dataset used for all our experimentations comes from a study of cross-experiment compiling microarray data of over 40.000 publicity available Affymetrix HG-U133Plus2 chip arrays (Torrente *et al.*, 2016). These arrays were produced under different experimental protocols and concerned different types of cancer. After pre-processing, the dataset contains 54675 probes for 22309 samples whose 14749 (66,11%) are cancer and 7560 (33.89%) are non-cancer. The full dataset is available on the ArrayExpress database under the identifier E-MTAB-3732. We split it into a training set of 17847 examples (11799 cancer, 6048 non-cancer) and a test set of 4462 examples (2950 cancer, 1512 non-cancer). Note that the original proportions of cancer and non-cancer samples are preserving in each set. The training set is used to train predictive models and the test to evaluate their performances.

3.2 Performances and sensitivity analysis

Deep GONet model is learned from the training set using a standard procedure. Dropout layers with a ratio of 0.6 have been added after each hidden layer to reduce overfitting. The weights and biases are initialized with He initializer. The network parameters are optimized using adam with an adaptive learning rate of 0.001. Different values of the hyperparameter α, controlling the regularization term L_{GO}, are tested in the interval $[0, 10^1]$. The number of epochs of the training is set up to 600. All the experiments have been executed on a GPU RTX 2080Ti using Tensorflow 1.12. The accuracy of the model is estimated from the test set according to the value of α to investigate the impact of this hyperparameter on the performance of the model. To reduce the variability of the results coming from the random initialization of the model parameters, 10 models for each value of α are learned with different random seeds for the initialization of the parameters.

Our method is compared with classical fully-connected network using L_2 or L_1 regularization terms. These regularization terms apply a penalty on all the connections regardless of the type (GO or noGO). L_2 is the squared magnitude of the weights $L_2=\sum_{l=1}^{L}\|W^{(l)}\|_2^2$ and L_1 is the absolute value of the magnitude of the weights $L_1=\sum_{l=1}^{L}|W^{(l)}|$. These regularization terms are also controlled by a hyperparameter α. These models are trained and tested with the same procedure used for Deep GONet. In addition, a model without any regularization is tested for comparison at $\alpha=0$. Note that all these models use the same baseline

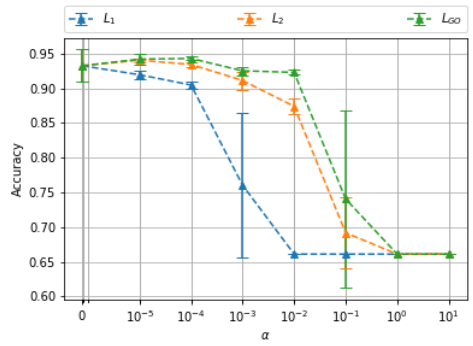

Fig. 2. Accuracy of the models according to α.

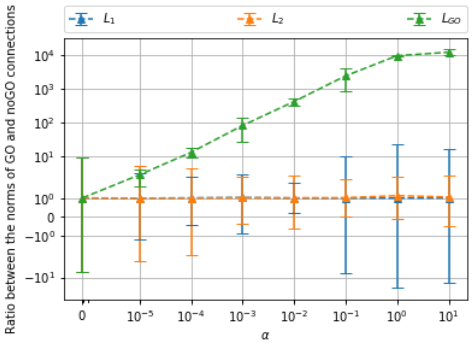

Fig. 3. Ratio between GO and noGO connections weights according to α.

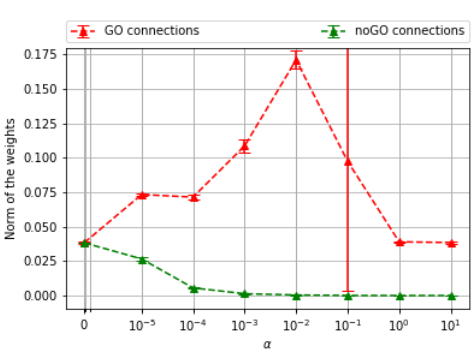

Fig. 4. Absolute-value norms of the GO and noGO connections from L_{GO} models according to α.

described in Fig. 1. For the results presented, we report the error bars evaluated on the test set.

Fig. 2 plots the average and the standard deviation of the accuracy of a model with a L_1, L_2, and L_{GO} penalty according to α. The three curves begin in the same point since $\alpha=0$ corresponds to a model without regularization. We can see that the model without regularization and, the one with L_{GO} and L_2 at $\alpha=10^{-5}$ achieve the best accuracy (0.945). Note that L_{GO} and L_2 outperform L_1. We also test classical machine learning methods on this dataset and obtain similar accuracy, 0.948 for SVM, 0.938 for XGboost, and 0.901 for Random Forest. These results show that our method obtains the same accuracy with the state-of-the-art algorithms, which are not self-explainable. For all models, the average accuracy decreases for high value of α. The accuracy drops to 0.66 which corresponds to the proportion of the majority class, meaning that the models learn nothing and associate all examples to the cancer class.

In this case, the regularization term takes too much importance relative to the cross-entropy. We note some special points, at $\alpha = 10^{-1}$ (for L_{GO} and L_2) and $\alpha = 10^{-3}$ (for L_1), with high variability. At these value of α, some models fail to learn with an accuracy of 0.66, whereas others succeed by reaching an accuracy around 0.9. That's why the average is between these two extremes.

Fig. 3 displays the ratio between the absolute-value norms of the GO (Eq. (2)) and noGO (Eq.(3)) connections, defined respectively as:

$$\frac{1}{L}\frac{1}{\#GO}\sum_{i=1}^{L}|W^{(l)} \otimes (C^{(l)})|, \tag{2}$$

$$\frac{1}{L}\frac{1}{\#noGO}\sum_{i=1}^{L}|W^{(l)} \otimes (1 - C^{(l)})|. \tag{3}$$

For L_2 and L_1, the ratio is stuck to 1 whatever the value of α. As expected, no distinction is made between the two types of connections. On the opposite, the ratio of the model with L_{GO} regularization increases along with the growth of α, and finally reaches its highest value of 10^4. For this model, Fig. 4 shows the average of the absolute-value norms of the GO (Eq. (2)) and noGO (Eq.(3)) connections. Note that the green line in Fig. 3 is obtained from the division of the red line by the green line. We can observe that the average norm of the GO connections remains between 10^{-2} and 10^{-1}. In contrast, the average norm of noGO connections decreases with α, following the accuracy trend. With $\alpha = 0$ and $\alpha = 10^{-5}$, the average norm of the noGO connections is very close to the one of GO connections. The ratio between the two norms, illustrated in Fig. 3, is below 10^1. From 10^{-4} to 10^1, the gap between the two norms becomes larger. The norm of noGO connections is converging almost to 0, leading to a ratio of 10^1 at $\alpha = 10^{-4}$ and the highest ratio of 10^4 at $\alpha = 10^1$. As a consequence, L_{GO} seems to penalize well the noGO connections with large value of α. However, the accuracy curves in Fig. 2 show that with a large value of α, the model is not able to learn anymore. It means that some noGO connections may be helpful for the prediction. In particular, the flexibility brought by the fully-connected architecture makes it possible. This advantage will be further inspected in subsection 3.4.

In summary, imposing a number of layers and neurons is not enough to make the model interpretable. An appropriate regularization term should be added to the loss function to constrain it along with biological knowledge. If the regularization term is not customized, the GO and noGO connections will be considered identically like with a L_2 or L_1 regularization. This results in a non-interpretable model without any prior knowledge. Our model Deep GONet reaches similar prediction performances than the state of the art, in both (i) penalizing properly the noGO connections, and (ii) privileging enough the GO connections to let the major information flow by them.

The models at $\alpha = 10^{-2}$ achieve an average accuracy around 0.92 and an average ratio of 10^3. Since it is a good trade-off between noGO connections penalization and accuracy, we decided to analyze in detail one of the models learned with $\alpha = 10^{-2}$ in the rest of this paper. The chosen model has an accuracy of 0.925 and a ratio of 10^3.

3.3 Analysis of the Deep GONet architecture

Table 1 presents in detail the architecture of the network selected in Subsection 3.2. The first two rows summarize the corresponding levels from GO-BP and the number of neurons from the first hidden layer to the sixth one (see Fig. 1). The last two rows give for each hidden layer the number of incoming connections (GO and noGO) and the number of incoming GO connections. Note that the total number of connections plus the number of neurons constitute the number of parameters of the model. The number of connections decreases over the layers because the number of neurons by layer becomes smaller. This table shows that the large majority of the connections are noGO connections, only 0.05% are GO connections.

hidden layer	1	2	3	4	5	6
level GO-BP	7	6	5	4	3	2
#neurons	1574	1386	951	515	255	90
#connections	86M	2.2M	1.3M	490K	131K	23K
#GO connections	43504	1709	1585	1010	491	175

Table 1. Hidden layers of Deep GONet.

Fig. 5 displays for each layer the sorting of the incoming connections according to the absolute value of their weight. The incoming GO (resp. noGO) connections are colored in red (resp. green). We first note that the connection matrices are very sparse, few connections have their weight significantly different from 0. This means that the gene expression is not uniformly propagated through the entire network and only a small part of the network is useful for the prediction. For all hidden layers, most of the GO connections are ranked before the noGO connections. Some of the GO connections can have a very high weight (around the 10^0). The high-weighted input GO connections of a neuron promote the activation of its corresponding function. The value of the noGO connections is close to 0 as expected by the application of the L_{GO} penalization. Some GO connections are ordered at the bottom of the rank. For example, the 43505th GO connection of the first layer is ranked 33041190th. The GO connections, which do not seem to be useful for the network, get a very low value (7.10^{-6} for our example). On the opposite, despite the application of the L_{GO} penalty on noGO connections, few of them have higher weight than GO connections as illustrated in the figure of the second hidden layer. These results show that the architecture of our model is very close to the GO-BP architecture since most of the weights of noGO connections are set to 0. The rare noGO connections with high weight are interesting. It represents links that the network has to build to compute accurate predictions. It would be interesting to investigate which GO terms or probes that have been connected by these noGO connections.

To continue the analysis of our network, we need to identify which neurons and connections are used to compute the predictions. For this, we employ a gradient-based method, *Layerwise Relevance Propagation* (LRP) (Bach *et al.*, 2015; Montavon *et al.*, 2017). The aim of LRP is to retropropagate the output signal of one sample from the upper hidden layer to the input layer. A relevance score is assigned to each neuron i of a layer l, described as $R_i^{(l)}$. This score represents the proportion of the output signal passing through its outcoming connections. The relevance of a neuron represents its importance in the computation of the prediction. For each sample, we get a relevance profile by layer composed of neuron relevances. An analysis of the neuron relevance of each layer confirms the fact that only a small subset of neurons is important to compute a given prediction.

3.4 Biological significance of the neurons

In this subsection, we check that the neurons of our network actually represent their corresponding GO term, i.e., the activation of a given neuron represents the expression of the corresponding biological function. For that, we use the fact that each GO term in GO is associated with a set of probes. If a given neuron really represents its corresponding GO term, the set of probes associated with this GO term should activate the neuron more than any other set of probes. We propose a procedure based on this principle to test the biological significance of the neurons. Because of the lack of space, we detail in the following only the analysis of the first hidden layer. However, similar analyses have been applied to the other layers.

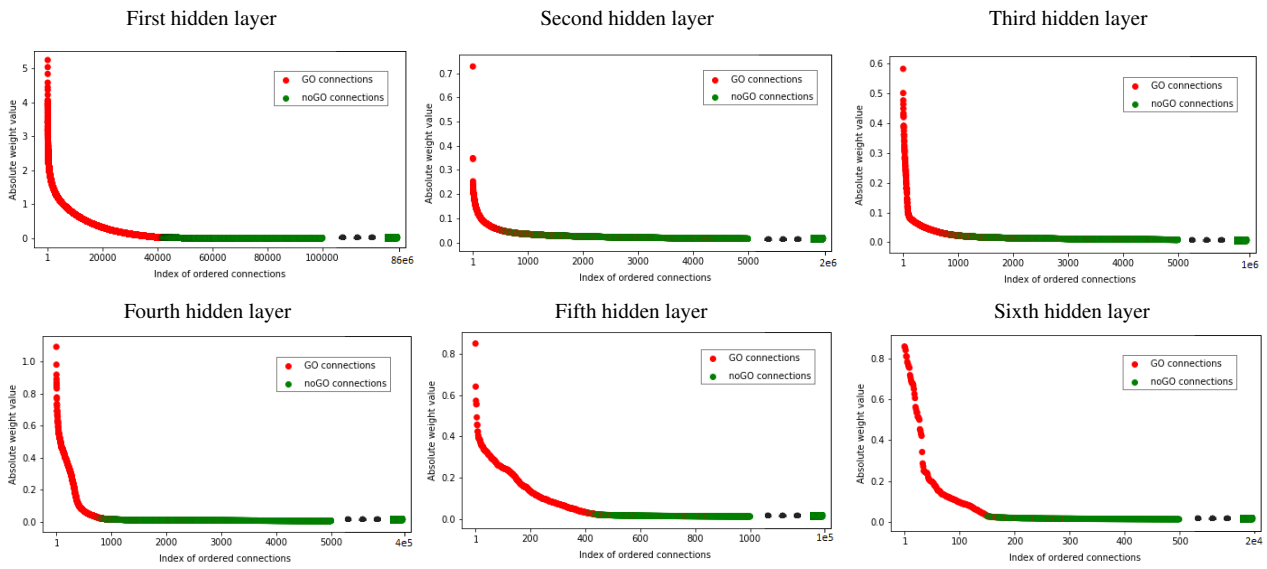

Fig. 5. Sorting of incoming connections from each layer according to their absolute weight value.

The first hidden layer contains 1574 neurons connected to the input layer. Each GO term is connected to a set of probes (median: 8, max: 1357, min: 1). Regarding this information, the target mask of a neuron is defined as follows:

- all the probes of the input layer, which are not connected to the GO term, are set to be 0;
- the values of the remaining probes in the set are unchanged.

In total, we have 1574 masks because none of the neurons has the same target mask. For every neuron, all these masks are applied to the input layer to identify whether the neuron is activated more by its target mask than the other masks. This can be measured by the rank of the target mask. The following procedure details how to get the rank of the target mask for each neuron in a layer l:

- Step 1: For each sample from the full test set, the activation $a_{i,m}^{(l)}$ of each neuron i for a given mask m is calculated. Note that there is no bias due to the length of the mask. Then, the average value of these activations is considered. For example, assuming that there are 3 neurons (3 masks) in the first hidden layer, for neuron 1, we obtain $\bar{a}_{1,1}^{(1)} = 0.7$, $\bar{a}_{1,2}^{(1)} = 0.9$ and $\bar{a}_{1,3}^{(1)} = 0.8$.
- Step 2: For each neuron, its activation values of all the masks are ordered in a decreasing way. Then, we have the rank $\bar{a}_{1,2}^{(1)}, \bar{a}_{1,3}^{(1)}, \bar{a}_{1,1}^{(1)}$. Suppose that the target mask of neuron 1 is mask 2, the rank of this mask is 1. Therefore, the neuron embodies the corresponding GO term.

To analyze the ranks of the target masks, we evaluate the relationship with the importance of the neurons by using LRP. The computation of $R_i^{(l)}$ follows the Step 1 and 2 without considering the masks. Since this method is applied sample by sample, the average of $R_i^{(l)}$ is also calculated for neuron i. For example, we acquire $\bar{R}_1^{(1)} = 1.9$, $\bar{R}_2^{(1)} = 2.5$ and $\bar{R}_3^{(1)} = 0.5$ respectively for the neurons $1, 2$ and 3. According to Step 2, the relevance scores are ordered in the following sequence $\bar{R}_2^{(1)}, \bar{R}_1^{(1)}, \bar{R}_3^{(1)}$. Then, based on this sequence, a rank to each neuron is attributed: neuron 2 gets the rank 1, and so on. Fig. 6 plots the rank of the target mask of the neurons along y-axis and their LRP relevance along x-axis. Note that the value of the ranks is up to the total number of neurons (i.e., 1574).

For the y-axis, a rank can get a NULL value or a discrete value in the range [1,16]. On the one hand, a NULL rank means that the activation

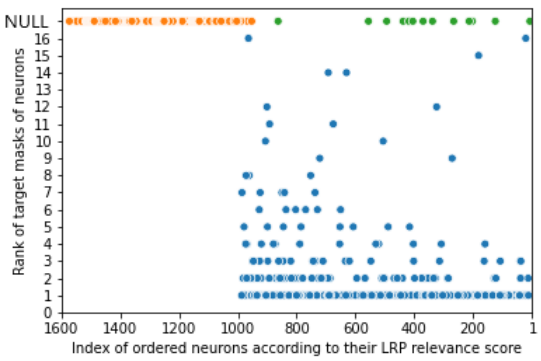

Fig. 6. Sorting of neurons from the first hidden layer according to the rank of their target mask (y-axis) and their LRP rank (x-axis).

of a neuron for its target mask is zero, which concerns 603 neurons (i.e., 38.31% of the 1574 neurons). Specifically, in total 591 neurons have a zero activation value for any mask and generally a LRP rank above the 1000-th order (colored in orange in Fig. 6). And the rest 12 neurons are activated by at least one another mask, and their LRP rank is below 1000-th order (colored in green in Fig. 6). On the other hand, there exist 971 neurons (61.69%) that have a positive activation for their target mask and show higher ranks, below order 1000. Among the 971 neurons, the target masks of 850 neurons rank 1, the other 121 neurons rank between 2 and 16. In conclusion, most of the neurons, which contribute highly to the prediction (LRP rank below order 1000), are well ranked for their target mask. This means that the important neurons for the prediction mainly match with their corresponding GO term.

Concerning the neurons with a NULL rank for their target mask, the major part has low LRP relevance. These neurons are not important for the predictions, they will not be therefore used in the interpretation. The associated GO terms can be ignored. However, the few neurons that have a high LRP relevance and a NULL rank are much more interesting. For instance, the neuron associated with GO:0071644 (*negative regulation of chemokine (C-C motif) ligand 4 production*) has a LRP rank of order 15,

but it is not activated by its target mask. Its target mask is composed of 2 probes, linked by GO connections weighted 0.1 and 0.04 respectively. On the opposite, 890 of the 1000 first noGO connections from the input layer, which have the same value with GO connections of norm-1 0.01, point this neuron. Since these neurons are not activated by their target mask, we can not say that they are associated with their corresponding GO term. We note that a large part of the noGO connections with high weight is connected to these neurons. Moreover, these noGO connections connect mainly probes with no annotation in GO, i.e., probes that have noGO connections. We can assume that the network distorts these neurons from their primary use to propagate the information of probes without GO annotations via noGO connections. These neurons do not represent anymore their corresponding GO terms but an unknown biological information useful for the predictions.

4 Biological interpretation of the results

In this section, we show how to use Deep GONet to provide a relevant biological interpretation of the model and its predictions.

4.1 LRP Relevance analysis

In this subsection, we study the clustering of samples predicted as cancer according to their relevance profiles. The relevance profile shows which part of the network is used to compute the prediction. We consider 2200 out of 2950 cancer samples for whom the tissue type is known. For each sample, a relevance profile constituted of the relevance of all neurons is computed with LRP. For each layer, we define a relevance matrix of size (N, N_l) containing the relevance of all neurons of this layer for all samples, where N is the number of samples and N_l the number of neurons in layer l. From these relevance matrices, we perform hierarchical clustering using the average linkage and the euclidean distance. The dendrograms of each layer are plotted in Fig 7. The colors on the dendrogram represent the type of tissue of the samples. From the dendrogram of the first hidden layer, we see that the patients from the same tissues tend to be grouped into the same clusters. It is especially the case for bone (colored in orange), blood (colored in red), lymph node (colored in blue), and ovary (colored in cyan). Tissues of the same type tend to share the same relevance profiles, meaning that some neurons are dedicated to one tissue. This clustering according to the tissue is still present in layer two although it is less significant. From the third hidden layer, the clustering of samples from the same tissues becomes less clear. From layer four to six, the clusters contain samples from different tissues. This means that important neurons for the predictions are used for different tissues. A signature of cancer shared by all tissues has been learned in the last layers of the network.

In conclusion, according to the way our architecture has been structured, the lower hidden layers gather more specific neurons and are responsible to extract features particular to a type of tissue. The model being more general incrementally, the upper hidden layers are instead in charge of extracting common features to any tissue. In the last hidden layer, the existence of several clusters indicates that different important neurons are used to predict the same prediction since the signal from the input layer to the output layer propagates along different paths. Note that for a classical MLP with L2 regularization, the dendrograms show only clusters with samples from different tissues for all layers. The architecture of Deep GONet allows the construction of a representation of gene expression specific to the type of tissue in the first layer of the network.

4.2 Model interpretation: Breast cancer

In this subsection, we show how to provide a biological interpretation of the model for a given disease. We focus on breast cancer to identify the most influential GO terms for the prediction. Breast cancer has been chosen

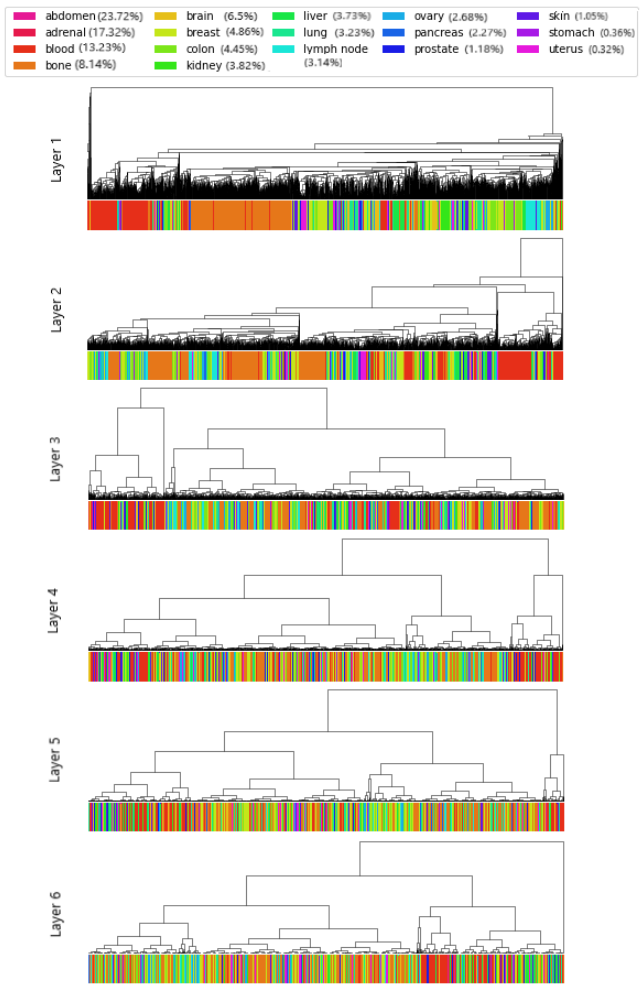

Fig. 7. Hierarchical clustering of test samples predicted as cancer based on their relevance profiles. Dendrograms are displayed by layer from the 1st hidden layer (top) to the 6th hidden layer (bottom).

because it is one of the most present cancer in our dataset. We compute the average LRP score of each neuron for all breast cancer samples. This relevance profile gives us which part of the network is used to predict breast cancer. By linking the important neurons (i.e., the highest LRP relevance) with their corresponding GO term, we can provide a biological interpretation of our neural network for breast cancer prediction.

Among all GO terms supporting the prediction of breast cancer in our network, some of them are known to be very related to this disease. In the first hidden layer, both terms GO:0030336 and GO:0030335 for negative and positive regulation of cell migration highlight the phenomenon of cancer cell invasion into surrounding tissues, which characterizes the beginning of tumor metastasis (Yamaguchi *et al.*, 2005). In the second hidden layer, GO:0016070 coding for RNA metabolic process refers to any events which occur in the life cycle of RNA, including their mutation. Some mutations can cause global alterations in the making of proteins and further lead to abnormal cells. More broadly, metabolic processes control the cell cycle by producing nutriments and energy. Any alteration of these processes can favor cancer (Hammoudi *et al.*, 2011). Two terms are related to intracellular signal activation, i.e., GO:1902531 (regulation of intracellular signal transduction) on the second hidden layer and GO:0035556 (intracellular signal transduction) on the third

hidden layer. Intracellular signal transduction is a chain of biochemical reactions transmitting signals from the cell surface to receptors of various components within the cell. It finally ends with a cellular response as a cell state change, i.e., cell growth and many other processes. It is part of cell communication. It was found that hyperactivity of these signal pathways can increase the proliferation of abnormal cells (Sever and Brugge, 2015). In the fourth hidden layer, GO:0042127 (regulation of cell proliferation) introduces the uncontrolled proliferation well known in tumor spread. In the same layer, we can notice the presence of a term, GO:0009725, dealing with response to hormone. Few types of cancer involving breast cancer are hormone-sensitive diseases. Estrogen and progesterone production can help them grow. In the fifth hidden layer, we have GO:0048512 encoding circadian behavior. Studies show that circadian rhythm can be affected by cancer (Fu and Kettner, 2013). Finally, on the sixth hidden layer, there is a new occurrence of a term related to proliferation, i.e., GO:0050673. It is specific to epithelial cell. An abnormal accumulation of epithelial cells often occurs with breast cancer (Dong-Le Bourhis *et al.*, 1997) but not only with this type of cancer (lung, colorectum…).

4.3 Use case of prediction interpretation

In this subsection, we show how to provide a biological interpretation of a given prediction. The objective is to propose a tool to the physicians and scientists that makes understandable the prediction computed by the model for a patient. This tool should point out the main biological functions used for the prediction computation and quantify their importance. To reach these objectives, we compute the LRP relevance of each neuron for a given prediction. Then, for each layer, the neurons are sorted according to their relevance and the most important ones are returned with their corresponding GO term and biological function.

Fig. 8 shows an example of the biological interpretation that we propose. In this example, we explain the prediction of the sample 24509 predicted as cancer at 99.99% by Deep GONet. For each layer, the five most important biological functions used are reported with their LRP relevance. In the first hidden layer, three terms (GO:0015031, GO:0006468, GO:0006412) are linked to protein activities that can highlight a high activity of protein disorder. In the second layer, both GO:0071420 and GO:1901258 reflect the immune activity response to cancer (Medina and Rivera, 2010; Chockalingam and Ghosh, 2014). The macrophage colony-stimulating factor is among one of the growth factors overexpressed in many tumors. Two additional terms related to protein (GO:0044257, GO:0010737) are present. Concerning GO:0010737, some dysregulations or mutations of the protein specie, i.e., kinase, contribute to all stages of cancer development (Bhullar *et al.*, 2018). In the third layer, the top-5 term (GO:0048864) can inform the production of cancer stem cells that have similar characteristics with normal stem cells. Between the fourth and fifth, different GO terms refer to membrane activity (GO:0071709, GO:0055085, GO:0044091). Many alterations of the membranes of tumor cells have been detected such as depolarisation (Yang and Brackenbury, 2013). GO:0050794 can point to the deregulation of cellular processes. Finally, concerning GO:0006739, studies show that the quantity of this molecule can be much higher in cancer cells (Ciccarese and Ciminale, 2017).

The comparison with two samples of different types of cancer (leukemia, lymphoma) shows that the interpretation of the last layer is very similar. At the final hidden layer, the ranking of these GO terms is almost the same no matter which cancer it is. In the first layers, the GO terms are very different. This confirms the conclusion that our network learned a specific representation of the samples in the first layer and a general representation of cancer shared by all samples in the last layers.

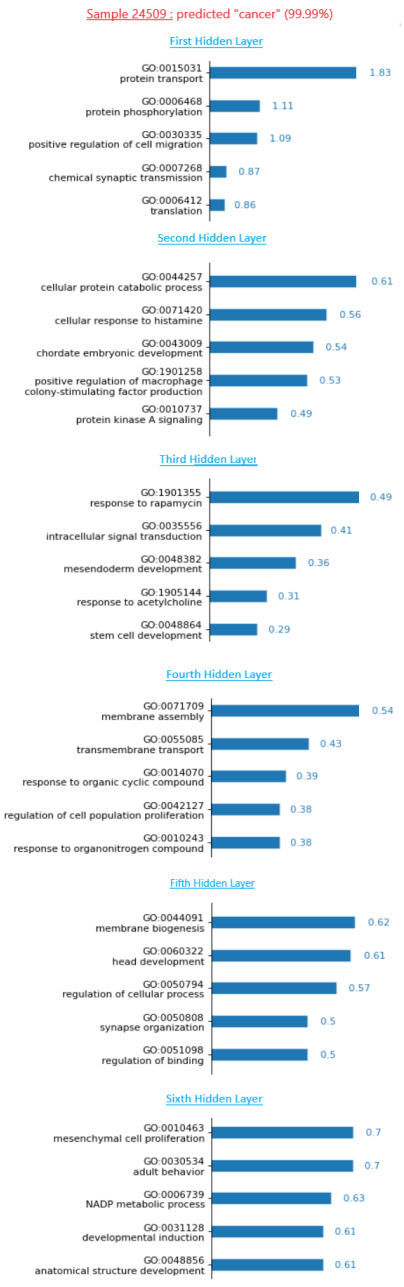

Fig. 8. Interpretation of the prediction of the sample 24509. The GO terms are ordered according to their relevance score for each layer.

5 Discussion and Conclusion

In this paper, we proposed, Deep GONet, a new self-explainable deep learning model for phenotype prediction based on gene expression. We demonstrate that its prediction performances are equivalent to classical deep learning and machine learning methods. The whole architecture of Deep GONet is interpretable and easy-understandable by biologists since it reflects the knowledge that they are used to employing. Each layer of Deep GONet corresponds to one level of GO and each neuron to a GO term. The addition of a customized regularization L_{GO} helps the model to better respect this knowledge by focusing on the real connections between the biological objects. The experiments presented on the cancer

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

detection show how to provide easily an interpretation of the model and its predictions, understandable by physicians and biologists. In this paper, the architecture of Deep GONet is based on GO-BP, but other ontologies can be implemented in the neural network with the same approach.

We point out that the goal of the interpretation is to explain how the model works and not how the biology works. Sometimes, there is no obvious relation between the biological functions, returned by the interpretation, and predicted phenotype. It is the case for half of the functions returned in the use case in subsection 4.3. This does not necessarily mean that the predictions are not reliable. We remind that a model looks for correlations between the output and the input and not for causalities. When a function, which seems not related to the phenotype, is returned, it is possible that this function either has an indirect correlation or is linked by an unknown causality relation with the phenotype. However, the more biological functions returned by the interpretation are coherent with the phenotype, the more we can trust the model predictions. If the most part of the interpretation is incoherent with the current biological knowledge, the reliability of the model should be interrogated. The model may overfit or be mislead by a bias in the training set.

Although the model interpretation is not a tool for biological discovery, some parts of our neural network could be investigated in this way. We refer, in particular, on the high-weighted noGO connections and neurons diverted from their GO term. These elements connect to the network the probes that have not annotations. It could be interesting to understand why these probes have been used for prediction, they should be related to the phenotype. We could also investigate how the expression of these probes is combined into the hidden layer. The probes connected to the same neuron could have close biological functions related to the predicted phenotype.

In future works, we plan to improve Deep GONet by adding neurons to deal with the probes without GO annotations. We will also turn the architecture into a graph neural network in order to represent the whole GO and not only some levels. We will finally investigate the links between the activation of a neuron and the expression of the corresponding biological function.

References

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, **6**, 52138–52160.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2019). Gradient-based attribution methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 169–191. Springer.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, **10**, 1–46.

Bhullar, K. S., Lagarón, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P., and Rupasinghe, H. P. V. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer*, **17**, 48.

Chockalingam, S. and Ghosh, S. S. (2014). Macrophage colony-stimulating factor and cancer: a review. *Tumor Biology*, **35**(11), 10635–10644.

Ciccarese, F. and Ciminale, V. (2017). Escaping death: mitochondrial redox homeostasis in cancer cells. *Frontiers in oncology*, **7**, 117.

Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, **32**(suppl_1), D258–D261.

Danaee, P., Ghaeini, R., and Hendrix, D. (2017). A deep learning approach for cancer detection and relevant gene identification. *Pacific Symposium on Biocomputing 2017*, **22**, 219–229.

Dong-Le Bourhis, X., Berthois, Y., Millot, G., Degeorges, A., Sylvi, M., Martin, P. M., and Calvo, F. (1997). Effect of stromal and epithelial cells derived from normal and tumorous breast tissue on the proliferation of human breast cancer cell lines in co-culture. *International journal of cancer*, **71**(1), 42–48.

Fakoor, R., Ladhak, F., Nazi, A., and Huber, M. (2013). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare*, volume 28.

Fu, L. and Kettner, N. M. (2013). The circadian clock in cancer development and therapy. *Progress in molecular biology and translational science*, **119**, 221–282.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Guo, W., Xu, Y., and Feng, X. (2017). DeepMetabolism: A deep learning system to predict phenotype from genome sequencing. *eprint arXiv:1705.03094*.

Hammoudi, N., Ahmed, K. B. R., Garcia-Prieto, C., and Huang, P. (2011). Metabolic alterations in cancer cells and therapeutic implications. *Chinese journal of cancer*, **30**(8), 508–525.

Hanczar, B., Henriette, M., Ratovomanana, T., and Zehraoui, F. (2018). Phenotypes prediction from gene expression data with deep multilayer perceptron and unsupervised pre-training. *International Journal of Bioscience, Biochemistry and Bioinformatics*, **8**, 125–131.

Hao, J., Kim, Y., Kim, T.-K., and Kang, M. (2018). PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics*, **19**(1), 510.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kang, T., Ding, W., Zhang, L., Ziemek, D., and Zarringhalam, K. (2017). A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. *BMC Bioinformatics*, **18**(1), 565.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, **13**, 8–17.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Medina, V. A. and Rivera, E. S. (2010). Histamine receptors and cancer pharmacology. *British journal of pharmacology*, **161**(4), 755–767.

Melis, D. A. and Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784.

Montavon, G., Samek, W., and Müller, K.-R. (2017). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, **73**, 1–15.

Peng, J., Wang, X., and Shang, X. (2019). Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. *BMC Bioinformatics*, **20**(8), 284.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144. Association for Computing Machinery.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1**(5), 206–215.

Sever, R. and Brugge, J. S. (2015). Signal Transduction in Cancer. *Cold Spring Harbor perspectives in medicine*, **5**(4), a006098.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR.org.

Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000). String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, **28**(18), 3442–3444.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR.org.

Torrente, A., Lukk, M., Xue, V., Parkinson, H., Rung, J., and Brazma, A. (2016). Identification of Cancer Related Genes Using a Comprehensive Map of Human Gene Expression. *PLOS ONE*, **11**(6), e0157484.

Yamaguchi, H., Wyckoff, J., and Condeelis, J. (2005). Cell migration in tumors. *Current opinion in cell biology*, **17**(5), 559–564.

Yang, M. and Brackenbury, W. J. (2013). Membrane potential and cancer progression. *Frontiers in physiology*, **4**, 185.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nature genetics*, **51**(1), 12–18.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture