Unveiling new disease, pathway, and gene associations via multi-scale neural networks

Thomas Gaudelet 1, Noël Malod-Dognin 2, Jon Sánchez-Valle 2, Vera Pancaldi 2,3,4, Alfonso Valencia 2,5,6 and Nataša Pržulj 2,5,∗

1 Department of Computer Science, University College London, London, WC1E 6BT
2 Department of Life Sciences, Barcelona Supercomputing Center (BSC), Barcelona, 08034 Spain
3 Centre de Recherches en Cancérologie de Toulouse (CRCT), UMR1037 Inserm, ERL5294 CNRS, 31037 Toulouse, France
4 University Paul Sabatier III, Toulouse, France
5 ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
6 Coordination Node. Spanish National Bioinformatics Institute, ELIXIR-Spain (INB, ELIXIR-ES), Spain
∗natasha@bsc.es.

Abstract

Diseases involve complex modifications to the cellular machinery. The gene expression profile of the affected cells contains characteristic patterns linked to a disease. Hence, new biological knowledge about a disease can be extracted from these profiles, improving our ability to diagnose and assess disease risks. This knowledge can be used for drug re-purposing, or by physicians to evaluate a patient’s condition and co-morbidity risk.

Here, we consider differential gene expressions obtained by microarray technology for patients diagnosed with various diseases. Based on these data and cellular multi-scale organization, we aim at uncovering disease–disease, disease–gene and disease–pathway associations. We propose a neural network with structure based on the multi-scale organization of proteins in a cell into biological pathways. We show that this model is able to correctly predict the diagnosis for the majority of patients. Through the analysis of the trained model, we predict disease–disease, disease–pathway, and disease–gene associations and validate the predictions by comparisons to known interactions and literature search, proposing putative explanations for the predictions.

Introduction

A disease is often described by symptoms and affected tissues. However, to give a definite diagnosis, physicians often need to analyse patient samples (e.g., blood samples, or biopsies) for characteristic disease indicators, commonly referred to as disease biomarkers. These may include disregulated genes, or pathways [16, 15]. Taking into consideration the history of a patient’s past and present conditions identifying genetic predispositions, as well as considering associations between diseases, aid in achieving accurate diagnostics and treatments [27]. By also taking advantage of the increasing availability of large scale molecular data, precision medicine aims at improving the understanding of the molecular base of diseases on individual basis, as well as the relationships between different conditions [19, 33].

The benefits from such work are multiple and include drug re-purposing and identification of new disease biomarkers to improve treatments and diagnoses.

Many studies have investigated disease–gene and disease–pathway associations with the objective of improving diagnoses [2, 24, 54, 26]. For instance, Zhao et al. [54] propose a ranking of disease genes based on gene expression and protein interactions using Katz-centrality. Hong et al. [26] design a tool that identifies significantly disrupted pathways by comparing patient gene expression against controls collected from other experiments. Cogswell et al. [9] identify putative gene and pathway biomarkers through change in miRNA in Alzheimer’s disease. In specific cancers, Abeel et al. [2] use support vector machines and ensemble feature selection methods to select putative gene biomarkers. Ciucci et al. [8] developed a general purpose algorithm based on the analysis of PCA loadings that can be used to identify genes that discriminate between conditions from expression data.

A key issue is that most of these studies consider diseases in isolation, i.e. comparing patients having a disease of interest to healthy individuals while the predicted biomarkers could be shared between various diseases. This limits the discriminative potential of such studies for accurate diagnoses. Indeed, network medicine has shown that diseases can share significant molecular background, as evidenced by numerous studies based on patient historical records [24, 27, 28], biological knowledge of the diseases [19, 33, 22], or patient gene expression profiles [44]. For instance, Goh et al. [19] build a disease network, which connects diseases that share at least one gene which when mutated
is linked to both conditions. Lee et al. [33] construct a disease network of metabolic diseases, connecting pairs of
diseases for which associated mutated enzymes catalyze adjacent metabolic reactions. Hidalgo et al. [25] take a
different approach by building a disease network based on disease co-morbidities, i.e. two diseases are connected if
they tend to co-occur significantly in the patient populations. They used Medicare records of elderly patients to build
the network. He et al. [22] propose PCID (Predicting Comorbidity by Integrating Data), an approach to predict
disease co-morbidities by aggregating disease similarity scores derived from different data including protein–protein
interactions (PPIs), pathways, and functional annotations.

Sánchez-Valle et al. [44] define a disease network, named the Disease Molecular Similarity Network (DMSN) based
on patient’s differential expression profiles. In their study, the DMSN is generated using positive and negative relative
molecular similarities (RR) to measure disease similarity and dissimilarity, respectively, that is then interpreted as
an estimate of risk. First, a patient-patient similarity network is generated based on the similarities of patient’s
differential expression profiles. Next, using the relative similarity score, diseases are related to each other. The
resulting network is directed and each edge is associated to a positive or negative label indicating either an increased
or decreased risk of developing the target disease if the patient has the source disease. The underlying assumption
is that having a given disease can increase the risk of developing a disease characterized by a similar gene expression
profile.

In these various approaches, a key issue is that either a single data source is used, such as disease–gene mutational
data [19], or no new biological knowledge about a specific disease could be derived from the results (e.g., PCID [22]).

In this work, we propose an integrative framework based on artificial neural networks (NN) to predict disease–
disease links, as well as disease–pathway and disease–gene associations. We train the model to predict patients’
diagnoses based on differential gene expression. The NN’s structure is designed to mimic the cellular multi-scale
functional organization by integrating gene–pathway annotations. This approach follows on from a body of work
aiming to build neural networks based on prior information defining the structure of the network [39]. For instance,
Ma et al. [34] build a neural network using the Gene Ontology [4] directed acyclic graph as a template for the
connections. The neural network, named DCell, is then trained to predict phenotype related to cellular fitness from
genotype data. The trained DCell predicts cellular growth almost as accurately as laboratory observations.

We show that our framework achieves good predicting performances on our dataset. By analysing the trained
NN, i.e. the underlying weight matrices, we show that we can extract biological knowledge relevant for each disease.
Specifically, we use the trained NN to predict novel disease–pathway and disease–gene links and from those predictions
we extract disease similarity score used to identify putative co-morbidities. We show that our predictions are
biologically relevant against established ground-truths and verify the top predictions through manual literature
curation ensuring that the sources do not use the same data, to mitigate the risk of argument circularity.

1 Material & Methods

1.1 Datasets

We use the original dataset of patient gene expressions used by [44] in which each patient is associated to a single
disease (see Supplement [3]). For each patient, we define a differential gene expression vector of size corresponding
to the number of genes and in which the $i^{th}$ entry is equal to 1, −1, or 0 depending on whether the $i^{th}$ gene is
significantly (with 5% cutoff) over-, under-, or normally expressed, respectively, for that patient (see Supplement [3]
for details).

Pathway annotations were collected from Reactome database [17] (accessed December 2018). Only the lowest
pathways in the hierarchy are considered to avoid dealing with pathway interactions (i.e. pathways containing other
pathways). Of those pathways, only the ones that have a Traceable Author Statement (TAS) are kept. In total, we
consider 1,708 pathway annotations.

The final dataset contains 4,788 samples (patients) diagnosed by one of 83 diseases (see Table [7]). In total, 20,525
genes have their expressions measured, but only a subset is used as input to our method described in the following
section, as we restrict ourselves to genes associated to at least one pathway, which leaves 9,247 genes.

1.2 Neural network based data–integration framework

We propose a neural network (NN) predicting a patient diagnosis based on differential gene expression. The structure
of the neural network is based on molecular organization, more specifically gene–pathway annotations downloaded
from Reactome (see Figure [1]). We integrate molecular organization data into our model to reflect the idea that
complex diseases, such as cancer, can be the results of the perturbations of groups of genes, as opposed to a single
Using Reactome data allows us to incorporate prior knowledge into our model in the form of biologically meaningful groupings of genes.

![Diagram of neural network architecture](image)

Figure 1: Example of neural network architecture. For the first layer, the connections are defined by biological information, i.e. a unit representing a gene is connected to all the biological pathways that the gene is involved in. We do not add any prior knowledge on the last layer, thus it is fully connected.

A neural network can be expressed as a series of matrix multiplications interleaved with non-linear function (See Supplement 3 for more details). Here, we use the softmax function as the last non-linear function of the NN (common choice for multiclass classification problems) and the hyperbolic tangent non-linear function for hidden layers to allow a hidden unit to have values lying in $[-1, 1]$ to represent up- and down-regulations. We use the classical cross-entropy function to define the loss function. Our neural network architecture has only one hidden layer capturing gene–pathway links. Hereafter, we refer to this model as GPD (for Gene–Pathway–Disease).

As reference model, we use a multiclass logistic regression (MLR; no hidden layer). MLR and GPD have a different number of parameters (or free weights): (573, 314) for MLR and (137, 838) for GPD. Note that, due to this imbalance, we do not expect GPD to outperform MLR in the diagnosis prediction task.

To train both GPD and MLR, we first perform a 10-fold cross-validation to fix the number of training epochs. To fix the number of training epochs, we compute the average number of epochs at which the test loss is the smallest across the runs (see Figure 3). As the dataset is imbalanced, we use stratification to split the data, ensuring that at least one patient per disease is in the test set. Using this number of epochs, we perform another 10-fold cross-validation to evaluate the performance of our models. We use the Adam optimizer with learning rate 0.01 and the layer weights are initialized to small values using the initialization process proposed by He et al. We investigated the addition of classical regularisation techniques – L1, L2, and dropout regularisation – as a mean to reduce the capacity of the model to overfit. We found that the performances were better without any regularisations (see Table 8).

The neural networks are implemented with Tensorflow.

1.3 Predicting disease–disease, disease–pathway, and disease–gene relationships

To identify associations between diseases and genes or pathways, we perform sensitivity analysis (see Supplement 3). The association score between disease $d$ and unit $u$ (representing a gene, or a pathway) is measured by the intensity of the local variation of the output unit associated with $d$ with respect to perturbation of $u$. Intuitively, this score measures how the prediction score of disease $d$ is affected by disregulation of a gene/pathway: we quantify the change in one disease score induced by a disregulation in the gene expression, or pathway activation. We test this scoring approach for the prediction of disease–gene and disease–pathway associations. In particular, we rank disease–gene and disease–pathway pairs based on this score and test if the score correlates with known associations through a Precision-Recall and Receiver Operating Characteristic (ROC) analysis. Score thresholds could be inferred from the
precision-recall curves. These thresholds would be uninformative in general since they are tied to specific runs and datasets. Hence, we focus on manual validation of the top scoring associations.

Based on this score, we represent each disease by a set formed by the $k_{\text{genes}}$ highest scoring genes and a set containing the $k_{\text{pathways}}$ highest scoring pathways. We then score disease–disease associations using the Jaccard Index of their sets. The Jaccard Index of two sets $S_1$ and $S_2$ is defined as $\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$, where $|\cdot|$ represents the cardinality of a set. Following on from similar approach used in DisGeNET [42], we interpret those associations as co–morbidities. The number of highest scoring pathways and genes considered is set to 150 and 300, respectively, as those numbers gave the best results.

2 Results & Discussion

2.1 Classification performances

To validate the relevance of our model, we verify that the classification performances are at least on par with competing methods: MLR, Random Forest (RF), Bernoulli Naive Bayes (nB), and Support Vector Machine (SVM) algorithms (we use the implementation available through the scikit-learning python package [41]). We perform 10-fold cross-validation for the algorithms to fix the hyperparameters (numbers of trees 100, smoothing parameter 0.001 and penalty parameter 100, respectively) and retain the best performing models in terms of cross-entropy loss (the objective function of the neural networks).

We evaluate performances by computing 3 different scores: cross-entropy loss (CEL), micro- and macro-average precision ($\text{Pre}_\mu$ and $\text{Pre}_M$, respectively). $\text{Pre}_\mu$ give a measure of the overall precision of each classifier, while $\text{Pre}_M$ gives an average of the precisions across the different classes (diseases). The details of each score can be found in Supplement 3.

| Algorithm | CEL     | $\text{Pre}_\mu$ | $\text{Pre}_M$ |
|-----------|---------|------------------|----------------|
| GPD       | 1.09 ± 0.06 | 0.80 ± 0.01    | 0.71 ± 0.02    |
| MLR       | **1.01 ± 0.07** | **0.84 ± 0.01** | **0.76 ± 0.01** |
| RF        | 1.56 ± 0.24 | 0.80 ± 0.01    | 0.70 ± 0.03    |
| nB        | 10.63 ± 0.55 | 0.66 ± 0.01    | 0.60 ± 0.02    |
| SVM       | 1.42 ± 0.04 | 0.72 ± 0.02    | 0.59 ± 0.02    |

Table 1: Performances of different classifiers in terms of cross-entropy loss (CEL), micro- and macro-average precisions ($\text{Pre}_\mu$ and $\text{Pre}_M$, respectively). Each score is computed across the 10-fold cross-validation and we provide the standard deviation. Bold scores highlight the best scores for each metric.

We observe that the neural networks; MLR and GPD, give better, or at least on-par, performances when compared to RF, nB, and SVM classifiers as measured by our three metrics (see Table 1). This observation justifies the relevance of our GPD model. The best model appears to be the multinomial logistic regression (MLR), which corresponds to the most complex neural network model in terms of the number of parameters (or degree of freedom), since MLR has $\sim 4$ times more parameters than GPD. This analysis shows that using biological knowledge to guide the structure of neural networks, in the limit of the models proposed, does not improve classification performance compared to the multiclass logistic regression (MLR) and only offers slight improvement when compared to a RF classifier (see Table 1). However, we show, in the following Sections, that the trained GPD models can be more successfully exploited than MLR to extract biological information. Note as well that the gene–pathway information on which GPD relies is both noisy and incomplete, as biological data often is, and that performances should improve as knowledge improves.

Hereafter, we consider for each model (GPD and MLR) the trained NN that gave the lowest cross-entropy loss during the 10-fold cross validation.

2.2 Our GPD model uncovers molecular mechanisms of diseases

To uncover molecular mechanisms of disease, i.e., genes and pathways that are associated to specific diseases, we extract predictions from MLR and GPD using the approach described in Section 1.3. We test the performance of our disease–pathway and disease–gene associations predictions by comparing against established databases. We investigate the top predictions of the GPD model through manual search of the literature.
Predicting disease–gene associations

For each model, we compute disease–gene association scores as described in Section 1.3 and we test the validity of our predictions against DisGeNET database [42]. We compare the entire set of predictions against two baselines: the Frequency of Differential Expression (FDE) and the approach introduced by Zhao et al. [54] for de novo disease–gene association prediction (Katz). Those methods are detailed in the Supplement 3.

We use precision–recall and ROC curves to evaluate the performance of our approach and compute the areas under the curves (see Table 2 and Figure 4). Interestingly, we observe that the FDE score is a poor predictor of disease–gene associations. We further observe that GPD is the best performing models for this task with Katz coming second. The relatively poor overall performances can be partially attributed to the incompleteness of the reported disease–gene associations in DisGeNET. To corroborate this hypothesis, we search the literature for support for the top 10 predicted disease–gene associations by the best performing model, GPD (see Table 5). Note that none of those associations are reported in DisGeNET.

| Model     | AUROC | AUPRE  |
|-----------|-------|--------|
| GPD       | 0.59  | 8.5e−3 |
| MLR       | 0.52  | 6.3e−3 |
| Katz      | 0.55  | 7.3e−3 |
| FDE       | 0.50  | 5.3e−3 |

Table 2: Performance in terms of area under the ROC (AUROC) and area under the precision–recall (AUPRE) for the prediction of disease–gene associations for each methods

| Disease                                | Gene     | Literature support |
|----------------------------------------|----------|--------------------|
| Asthma                                 | UBB      |                    |
| Schizophrenia                          | RHOA     | PMID:16402129      |
| Alzheimer’s disease                    | FGF23    | PMID:26674092      |
| Autistic disorder                      | FGF20    | PMID:19204725      |
| Prostate cancer                        | RPS27A   | PMID:15647830      |
| Amyotrophic lateral sclerosis          | PSMD13   |                    |
| Amyotrophic lateral sclerosis          | CASP3    | PMID:11715057      |
| Chronic obstructive pulmonary disease  | SKP1     | PMID:23713962      |
| Autistic disorder                      | PSMB2    |                    |
| Irritable bowel syndrome               | PSMA1    | PMID:28717845      |

Table 3: Top 10 disease–gene predicted by GPD.

We are able to find literature support for 70% of the top 10 predicted disease–gene associations (see Table 3). Furthermore, we find indications that some of our top-scoring, non-validated predictions could be relevant, such as the associations of asthma with UBB and amyotrophic lateral sclerosis (ALS) with PSMD13. Ubiquitin B (UBB) belongs to the ubiquitin-proteasome (UPS) and it is known that aberration in the UPS is responsible for inflammatory and autoimmune diseases such as asthma [53]. Moreover, ALS onsets occur typically after age 50 and manifest partially through muscle weakness. PSMD13 is linked to aging [6] and high expression of the gene has been found in skeletal muscle of athletes [31], suggesting that under-expression could be a sign of muscle weakness.

These results validate the relevance of our framework for de novo disease–gene association prediction and confirm the incompleteness of DisGeNET.

Predicting disease–pathway associations

For our GPD model, we compute disease–pathway association scores as described in Section 1.3 and we test the validity of our predictions by comparison with CTD database [13]. As a baseline, we consider disease–pathway scores corresponding to the average FDE (AFDE) of genes within the pathway for patients having the disease.

We evaluate the results as done previously for disease–gene associations (see Table 2 and Figure 5). We observe that GPD convincingly outperforms AFDE. The seemingly poor performances of both approaches can partially be attributed to the incompleteness of CTD database. To test this hypothesis, we search the literature for support for the top 10 disease–pathway associations predicted with our GPD (see Table 5). Note that none of these associations predicted are reported in CTD database.
| AUROC   | AUPRE  |
|---------|--------|
| 0.53    | 8e^{-2}|
| 0.47    | 6e^{-2}|

Table 4: Performance in terms of area under the ROC (AUROC) and area under the precision–recall (AUPRE) for the prediction of disease–pathway associations for each methods

| Disease Pathway R-HSA- | Literature support |
|-----------------------|--------------------|
| Autistic disorder     | 5653890            |
| Irritable bowel syndrome | 532668             | PMID:20338921 |
| Irritable bowel syndrome | 391906             | PMID:16835707 |
| Type 2 diabetes mellitus | 499943             | doi:10.2337/diabetes.51.2007.S363 |
| Asthma                | 391906             | PMID:8603274   |
| Schizophrenia         | 71288              | PMID:22465051  |
| Major depressive disorder | 8934903           | PMID:27063986  |
| Type 2 diabetes mellitus | 8939245            | PMID:19667185  |
| Schizophrenia         | 5683371            |
| Sjogren’s syndrome    | 389661             |

Table 5: Top 10 disease–pathway predictions derived from GPD.

We find literature support for 7 out of the top 10 predicted disease–pathway associations (see Table 5). Furthermore, we find indications that some of our top-scoring, non-validated predictions could be relevant, such as the association of autistic disorder with the lactose synthesis pathway (R-HSA-5653890) and the association of Schizophrenia with pathway R-HSA-5683371 linked to microphthalmia. The lactose synthesis pathway (R-HSA-5653890) contains three genes: LALBA, SLC2A1, and B4GALT1. All of those genes might be associated with autistic disorders. One patented method to detect autistic disorder (US20140349977A1) includes LALBA as one of the genes of interest. SLC2A1 mutation has been reported in patients diagnosed with autism [21]. And finally, B4GALT1 has been linked with developmental disorders [22]. The pathway R-HSA-5683371 is linked to the eye disease microphthalmia. It is known that schizophrenia is linked to eye abnormalities [21]. Among the 28 genes involved in that pathway, 12 have been linked to the disease in the literature (GOT2, PDHA1, DLD, GCSH, DLAT, PDHB, DAO, OGDH, DHTKD1, GNMT, DDO, PRODH2).

These results show the relevance of our framework for de novo disease–pathway associations prediction despite relatively low retrieval scores against the ground-truth.

2.3 Our GPD model predicts disease–disease relationships

We rank disease–disease pairs based on the score described in Section 1.3 and test our results against a high confidence co-morbidity disease network obtained from a large cohort study by Hidalgo et al. [25]. We compare our method against DMSN network [14], restricted to our set of diseases, and three alternatives baselines. For the first alternative, we compute disease–disease association score using our approach defined in Section 1.3 on the trained MLR network, representing each disease by the top 300 highest scoring genes (which gave the best results based on grid search). The final two baselines associate to each disease–disease pair a Jaccard Index score based on 1) the set of genes associated to each disease in DisGeNET [42] and 2) the set of pathways associated to each disease in CTD database [13]. The results of the comparison are presented using a precision–recall curve (see Figure 2).

We observe that our approach outperform convincingly the other approaches in the task of retrieving existing co-morbidity links between diseases with over 10% increase compared to DMSN and 30% improvement over DisGeNET in terms of area under the precision–recall curve (auprc). These results strongly support our methodology. The scoring based on disease–gene is performing better than disease–pathway, hence we investigate the top 10 scoring disease–disease associations derived from it (see Table 6).

We present and discuss below literature support for the predicted associations between the diseases.

Atrial fibrillation has been linked in the literature to thyroid disease [47] which is known to be co-morbid with vitiligo [11]. Atrial fibrillation and peripheral vascular disease are well known co-morbid conditions [43]. Alcoholism has been linked to the onset of some cancers notably implicating the transcription factor Nanog which itself has been linked to osteosarcoma [35]. Additionally, a drug used to treat alcoholism, Disulfiram, has recently been proposed as a potential treatment for osteosarcoma [10]. Those information put together suggest shared molecular background
Figure 2: Precision–recall (top) and ROC (bottom) curves of the test against the disease co-morbidity network built by Hidalgo et al. [25].

| Disease 1                  | Disease 2                  |
|----------------------------|----------------------------|
| Atrial fibrillation        | Vitiligo                   |
| Atrial fibrillation        | Peripheral vascular disease|
| Alcoholic hepatitis        | Osteosarcoma               |
| Rhabdoid cancer            | Medulloblastoma            |
| Cornelia de Lange syndrome | Vitiligo                   |
| Peripheral vascular disease| Vitiligo                   |
| Atrial fibrillation        | Osteosarcoma               |
| Leishmaniasis              | Alcoholic hepatitis        |
| Sotos syndrome             | Vitiligo                   |
| Follicular lymphoma        | Osteosarcoma               |

Table 6: Top 10 disease–disease links predicted using our approach based on the trained GPD.

for the two conditions. Rhabdoid cancer is a rare form of aggressive cancer affecting young children and with very poor prognostic, which makes it difficult to evaluate co-morbid conditions. However, rhabdoid cancer is frequently mistaken for medulloblastoma indicating some similarity [7]. We found no evidence in the literature supporting a connection between the rare Cornelia de Lange syndrome and Vitiligo. Some studies have observed significant co-morbidity between vitiligo and psoriasis and the combination of the two has been linked to cardiovascular diseases, which include peripheral vascular disease [3, 11]. Atrial fibrillation and cardiac complications have been observed as the result of osteosarcoma [49, 14]. Leishmaniasis and alcoholic hepatitis are an unlikely co-morbid connection since it would require a patient both to have been infected by parasites of the Leishmania type and have had excessive alcohol intake. However, both disease affects the liver and leishmaniasis has sometimes been misdiagnosed for cirrhosis [18] which suggests that the two diseases might share some similar molecular processes that we would be capturing here. A case of co-occurrence of Sotos syndrome and vitiligo has been reported in the medical literature [12]. It has been postulated that non-Hodgkin’s lymphoma (which include follicular lymphoma) and osteosarcoma share underlying mechanisms [39]. Additionally, miR-202 has been identified as a potential tumor suppressor for both conditions [50].

Through this analysis, we have shown that most predicted pairs have either been observed co-occurring or can be connected through underlying mechanisms, thus validating our approach.

3 Conclusions

In this study, we propose a multi-scale neural-network based framework that integrates gene expression data associated to diseases with gene–pathway information. Our integrative framework allows for simultaneously uncovering novel disease-disease associations and disease molecular mechanisms from patient gene expression profiles through
the analysis of trained neural networks. We show that GPD achieve good diagnosis prediction on our dataset showing the validity of our integrative process. Furthermore, we show that the associations predicted from the trained models are biologically meaningful and supported by the literature, thus validating our approach.

While our disease molecular mechanisms are supported by the current knowledge about these diseases, a next step would be to identify among the predicted genes and pathways suitable biomarkers and drug targets that could be used to improve diagnosis, prognosis, and treatment. We leave this for future work. Also, while our multi-scale NN framework integrates the hierarchical functional organization of a cell (from genes to biological pathways), our methodology can be extended to include any dataset pertaining to diseases of interest, e.g., uncovering molecular mechanisms of cancer from patient somatic mutation profiles or linking diseases to non-coding RNA.

Finally, while we focus on patient data with application to diseases, our methodology can be extended to integrate additional omics data with the objective of getting more biologically accurate models for the analyses of patients, tissues, and cells. Some further applications include studies of diseases linked to a specific tissue, studies of cell’s specialization, and any study that can benefit from the integration of the hierarchical functional organization of cells.

Funding

This work was supported by the European Research Council (ERC) Consolidator Grant 770827, UCL Computer Science, the Slovenian Research Agency project J1-8155, the Serbian Ministry of Education and Science Project III44006, the Prostate Project, the Fondation Toulouse Cancer Santé and Pierre Fabre Research Institute as part of the Chair of Bio-Informatics in Oncology of the CRCT, PhD Fellowship (BES-2016-077403) and the Spanish Ministry of Economics and Competitiveness (BFU2015-71241-R).

Supplement

Project webpage  
https://life.bsc.es/iconbi/MultiScaleNN/index.html

SI Datasets  The base data used in this project is collected from the work of Sanchez-Valle et al. [44]. It consists of multiple datasets of gene expressions captured by micro-array technology [16]. The datasets are downloaded from Gene Expression Omnibus (GEO, [5]) and ArrayExpress [30] databases. Each dataset contains measurements from healthy (controls) and affected (patients) subjects. For a given dataset, the measurements originate from bulk samples extracted from the same tissue in each subject. Not all datasets use the same tissue for measurements, as diseases do not necessarily affect the same tissue. Each patient is diagnosed with a single disease. For comparisons, the data is normalized by using the frozen robust multiarray procedure [37] to remove experimental bias. Furthermore, to remove tissue effects, each patient sample is normalised against all the control samples of its original dataset using the Limma method [48]. Up to this point, the data is identical to those used to derive the DMSN network [44]. Then we use the corrected p-values output by Limma to define for each patient a vector of size corresponding to the number of genes and in which the $i^{th}$ entry is equal to 1, −1, or 0 depending on whether the $i^{th}$ gene is significantly (with 5% cutoff) over-, under-, or normally expressed, respectively, for that patient. Additionally, we exclude patients that have no significantly deregulated genes, as we cannot learn anything from them.

The set of diseases is curated by hand for associations with Disease Ontology codes [45], standard ICD9 and ICD10 codes, MeSH terms and OMIM codes [38]. Some of the datasets come from studies investigating subtypes of diseases that are studied by projects linked to other datasets. Based on the number of patients in each study, those datasets were either merged with the more global disease, or the patients associated with the more global disease were dropped from the study. Specifically, we drop the global disease if the subtype has many more patients associated with it and merge otherwise. Finally, we exclude diseases that have less than 10 associated patients to capture disease heterogeneity in the final dataset and to have sufficient data for each disease to split in a training and testing set.

SI Model  A neural network can be expressed as a series of matrix multiplications interleaved with non-linear functions, formally the output $Y$ of a neural network with $n-1$ hidden layers can be written as

$$Y = f_n(W_nf_{n-1}(\ldots f_1(W_1X)))$$

where $X$ represents the input data, $W_i$ the weights of layer $i$, and $f_i(\cdot)$ the non-linear function applied to the output of the $i^{th}$ layer. The optimization problem can be written as the minimization of the loss function $L = g(\hat{Y}, Y)$.
where \( \hat{Y} \) is the objective, or ground truth, and \( g(\cdot) \) is a predefined function. Multinomial logistic regression (MLR) and the proposed GDP architecture can be written as

\[
\begin{align*}
Y^1 &= s(W^1 X) \\
Y^2 &= s(W^2_2 \tanh(W^2_1 X))
\end{align*}
\]

(1)

(2)

where \( s \) is the softmax function, typically used for multiclass classification problems and \( \tanh \) denotes the hyperbolic tangent. Matrices \( X \) and \( Y \) represent our data. Each column of \( X \) corresponds to the differential gene expression of a patient, and each column of \( Y \) corresponds to a patient’s diagnosis (the prediction of which is the objective of the framework). \( W^1 \in \mathbb{R}^{n_d \times n_g} \) and \( W^2_1 \in \mathbb{R}^{n_d \times n_p} \) correspond to fully-connected layers. The layer corresponding to \( W^2_1 \in \mathbb{R}^{n_d \times n_p} \) represents biological pathway membership of the genes, i.e. the trainable weights of the matrix correspond only to entries \((i,j)\) where gene \( j \) is part of the \( i^{th} \) protein complex.

**SI Methods** Formally, the local variations \( \delta f \) of a single-argument function \( f \) due to a change \( \delta x = x - x_0 \) in input can be approximated with the first order Taylor expansion as

\[
\delta f(x) = \frac{df}{dx} (x_0) \delta x + O(x^2).
\]

Thus, the magnitude of the local variations of \( f \) with respect to perturbation \( \delta x \) from \( x_0 \) is given by \( \left| \frac{df}{dx} (x_0) \right| \). Based on this approximation, we extract from each neural network a score between an entity represented by a unit of the neural network (e.g. a pathway, or a gene) and each disease (output unit). Specifically, for a neural network \( NN \), we denote \( nn_i : [0,1]^n_i \rightarrow \mathbb{R}^{n_o} \) the function corresponding to the operation of a neural network \( NN \) from the \( n_i \) outputs of layer \( i \) to the final \( n_o \) logits of the neural network, i.e. scores before softmax. E.g., for GPD, we have \( nn_j^2(x) = W^2_2 \tanh(W^2_1 x) \). Then, the association score \( s_{i,j,k} \) between the \( j^{th} \) output unit of layer \( i \), denoted \( u_j^i \), of neural network \( NN \), and disease \( k \) is given by

\[
s_{i,j,k} = \left| \frac{\partial nn_i}{\partial u_j^i}(x_0) \right|_k,
\]

where the reference point is chosen as the null vector, \( x_0 = 0 \), which corresponds implicitly to a healthy state in our formulation.

**SI Metrics** The cross-entropy loss (CLE) of a classifier is defined as

\[
CLE = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} -y_{ij} \log(\hat{y}_{ij}),
\]

where \( m \) represents the number of samples (patients), \( n \) the number of classes (diseases), \( y_{ij} \) indicates if patient \( i \) is diagnosed with disease \( j \) \((1 \text{ if true } 0 \text{ otherwise})\), and \( \hat{y}_{ij} \) is the \( j^{th} \) output value of the classifier for patient \( i \).

A relatively small CEL means that the output probability distribution of a classifier is closer to the deterministic one-hot encoding of the true labeling, i.e. the classifier gives a higher probability to the true class and a very small probability to the other classes.

The micro-averaged precision (Pre\(_\mu\)) of a classifier gives a measure of the overall precision of the classifier and is defined as

\[
\text{Pre}_\mu = \frac{tp}{m},
\]

where \( tp \) corresponds to the number of accurately classified patients and \( m \) represents the number of patients.

The macro-averaged precision (Pre\(_M\)) of a classifier gives an average of the precision across the different classes (diseases) and is defined as

\[
\text{Pre}_M = \frac{1}{n} \sum_{i=1}^{n} \frac{tp_i}{m_i},
\]

where \( tp_i \) corresponds to the number of accurately classified patients for disease \( i \) and \( m_i \) represents the number of patients diagnosed with the same disease. Pre\(_M\) can be more informative than Pre\(_\mu\) when considering the problem with class imbalance.
**SI Baselines**  The Frequency of Differential Expression (FDE) score of a disease–gene association corresponds to how frequently that gene is consistently differentially expressed in patients having the disease, i.e., for disease \( d \) and gene \( g \), the association score, \( s_{dg} \), is given by

\[
s_{dg} = \frac{1}{|\mathcal{P}_d|} \sum_{p \in \mathcal{P}_d} X_{gp},
\]  

where we amalgamate entities (disease, gene, and patient) with their indices, \( \mathcal{P}_d \) denotes the set of patients having disease \( d \), and \( X \) corresponds to the data matrix introduced in Methods.

The Katz method uses disease specific Protein–Protein Interaction (PPI) network, where each node of a standard PPI network is associated to a score (here the FDE of each gene for the disease considered). The authors then use Katz-centrality on each disease PPI network to extract a final score for each disease–gene association (here we use the absolute value). The higher the score, the higher the association is expected to be true. We download the PPI data from BioGRID [40] and IID [32] and create our PPI network from the union of both databases restricted to our set of genes. Finally, we perform a grid-search to identify the best parameters for the model by trying to maximise the area under the precision-recall curve metric.

| Disease Name                                      | Patients Count | Disease Name                        | Patients Count |
|--------------------------------------------------|----------------|-------------------------------------|----------------|
Table 7: Cohort size for each disease in the dataset.

| Disease               | Size |
|-----------------------|------|
| medulloblastoma       | 41   |
| leishmaniasis         | 11   |
| polycythemia vera     | 41   |
| follicular lymphoma   | 10   |
| atopic dermatitis     | 40   |
| myelofibrosis         | 10   |
| trachoma              | 39   |
| leigh disease         | 10   |
| rosacea               | 38   |

Table 8: Results of cross-validation to fix regularisation hyperparameters (L1-, L2-, or dropout regularisations). The scores correspond to cross-entropy loss. The best results are obtained with no regularisation (score in bold).

| hyperparameter | 10  | 0.1 | 0.01 | 0.001 | 0    |
|----------------|-----|-----|------|-------|------|
| L1-regularization | $2026 \pm 4.3$ | $24.5 \pm 0.046$ | $6.39 \pm 0.003$ | $2.71 \pm 0.047$ | $\mathbf{1.09 \pm 0.066}$ |
| L2-regularization | $4.42 \pm 4.8e^{-7}$ | $4.38 \pm 0.012$ | $3.48 \pm 0.032$ | $2.02 \pm 0.026$ | / |
| dropout ratio     | 0.25 | 0.5 | 0.75 | 0.9   | 0    |
| dropout           | $1.14 \pm 0.121$ | $1.10 \pm 0.130$ | $1.12 \pm 0.086$ | $1.10 \pm 0.081$ | / |

Figure 3: Train and test loss curve for the MLR model (left) and GDP model (right) with respect to the number of epochs during the cross-validation. The vertical black line indicates the number of epochs which give the lowest loss.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pages 265–283, Savannah, GA, 2016. USENIX Association.

[2] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3):392–398, 2009.

[3] M Arunachalam, F Dragoni, R Colucci, S Berti, E Crocetti, M Galeone, R Conti, and S Moretti. Non-segmental vitiligo and psoriasis comorbidity–a case-control study in italian patients. Journal of the European Academy of Dermatology and Venereology, 28(4):433–437, 2014.
Figure 4: Precision-recall curve and ROC curve of our predictions for disease–genes associations.

Figure 5: Precision-recall curve and ROC curve of our disease–pathways associations predictions.
[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25, 2000.

[5] Tanya Barrett, Stephen E. Willhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashovskiy, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[6] D Bellizzi, S Dato, P Cavalcante, G Covello, F Di Cianni, G Passarino, G Rose, and G De Benedictis. Characterization of a bidirectional promoter shared between two human genes related to aging: Sirt3 and psmd13. *Genomics*, 89(1):143–150, 2007.

[7] Peter C Burger, IT Yu, Tarik Tihan, Henry S Friedman, Douglas R Strother, James L Kepner, Patricia K Duffner, Larry E Kun, and Elizabeth J Perlman. Atypical teratoid/rhabdoid tumor of the central nervous system: a highly malignant tumor of infancy and childhood frequently mistaken for medulloblastoma: a pediatric oncology group study. *The American Journal of Surgical Pathology*, 22(9):1083–1092, 1998.

[8] Sara Ciucci, Yan Ge, Claudio Durán, Alessandra Palladini, Víctor Jiménez-Jiménez, Luisa María Martínez-Sánchez, Yuting Wang, Susanne Sales, Andrej Shevchenko, Steven W Poser, et al. Enlightening discriminative network functional modules behind principal component analysis separation in differential-omic science studies. *Scientific Reports*, 7:43946, 2017.

[9] John P Cogswell, James Ward, Ian A Taylor, Michelle Waters, Yuming Shi, Brian Cannon, Kevin Kelner, Jon Kemppainen, David Brown, Caifu Chen, et al. Identification of mirna changes in alzheimer’s disease brain and csf yields putative biomarkers and insights into disease pathways. *Journal of Alzheimer’s Disease*, 14(1):27–41, 2008.

[10] Jared Anthony Crasto, Mitchell Stephen Fourman, Alejandro Morales-Restrepo, Adel Mahjoub, Jonathan Brendan Mandell, Kavita Ramnath, Jessica C Tebets, Rebecca J Watters, and Kurt Richard Weiss. Disulfiram reduces metastatic osteosarcoma tumor burden in an immunocompetent balb/c or-thotopic mouse model. *Onco-target*, 9(53):30163, 2018.

[11] Aisha M Dahir and Simon F Thomsen. Comorbidities in vitiligo: comprehensive review. *International Journal of Dermatology*, 57(10):1157–1164, 2018.

[12] Per Dahlqvist, Rupert Spencer, Pedro Marques, Mary N Dang, Camilla AM Glad, Gudmundur Johannsson, and Mártí Korbonits. Pseudocromegaly: a differential diagnostic problem for acromegaly with a genetic solution. *Journal of the Endocrine Society*, 1(8):1104–1109, 2017.

[13] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2019. *Nucleic Acids Research*, page gky868, 2018.

[14] Tomotaka Dohi, Hirotoshi Ohnura, Hiroyuki Daida, and Atsushi Amano. Progressive visceral leishmaniasis misdiagnosed as cirrhosis of the liver: a case report. *European Journal of Cardio-thoracic Surgery*, 35(3):544–546, 2009.

[15] Joel T Dudley, Robert Tibshirani, Tarangini Deshpande, and Atul J Butte. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology*, 5(1):307, 2009.

[16] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Ágnar Helgason, G Bragi Walters, Steinnunn Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423, 2008.

[17] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D’Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 2018.

[18] Lydia Giannitrapani, Maurizio Soresi, Emanuele La Spada, Claudio Tripodo, and Giuseppe Montalto. Progressive visceral leishmaniasis misdiagnosed as cirrhosis of the liver: a case report. *Journal of Medical Case Reports*, 3(1):7265, 2009.
[19] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[20] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press Cambridge, 2016.

[21] Natsuko Hashimoto, Kuriko Kagitani-Shimoto, Norio Sakai, Takanobu Otomo, Koji Tominaga, Shin Nabatame, Yukiko Mogami, Yukitoshi Takahashi, Katsumi Imai, Keiko Yanagihara, et al. Slc2a1 gene analysis of japanese patients with glucose transporter 1 deficiency syndrome. *Journal of Human Genetics*, 56(12):846, 2011.

[22] Feng He, Guanghui Zhu, Yin-Ying Wang, Xing-Ming Zhao, and De-Shuang Huang. Pcid: A novel approach for predicting disease comorbidity by integrating multi-scale data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 14(3):678–686, 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[24] Zengyou He and Weichuan Yu. Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, 34(4):215–225, 2010.

[25] César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353, 2009.

[26] Guini Hong, Hongdong Li, Jiahui Zhang, Qingzhou Guan, Rou Chen, and Zheng Guo. Identifying disease-associated pathways in one-phenotype data based on reversal gene expression orderings. *Scientific Reports*, 7(1):1348, 2017.

[27] Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5:4022, 2014.

[28] Venkateshan Kannan, Fredrik Swartz, Narsis A Kiani, Gilad Silberberg, Giorgos Tsipras, David Gomez-Cabrero, Kristina Alexanderson, and Jesper Tegnér. Conditional disease development extracted from longitudinal health care cohort data using layered network construction. *Scientific Reports*, 6:26170, 2016.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y. Amy Tang, Eleanor Williams, Mirosław Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, Karyn Megy, Ekaterina Pilicheva, Gabriella Rustici, Andrew Tikhonov, Helen Parkinson, Robert Petryszak, Ugis Sarkans, and Alvis Brazma. Arrayexpress update—simplifying data submissions. *Nucleic Acids Research*, 43(D1):D1113–D1116, 2015.

[31] Erika Koltai, Zoltan Bori, Peter Osath, Ferenc Ihasz, Szabolics Peter, Geza Toth, Hans Degens, Jörn Rittweger, Istvan Boldogh, and Zsolt Radak. Master athletes have higher mir-7, sirt3 and sod2 expression in skeletal muscle than age-matched sedentary controls. *Redox Biology*, 19:46–51, 2018.

[32] Max Kotlyar, Chiara Pastrello, Nicholas Sheahan, and Igor Jurisica. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research*, 44(D1):D536–D541, 2015.

[33] D-S Lee, Juyong Park, KA Kay, Nicholas A Christakis, ZN Oltvai, and A-L Barabási. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 2008.

[34] Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290, 2018.

[35] Keigo Machida, Chia-Lin Chen, Jian-Chang Liu, Claudine Kashiwabara, Douglas Feldman, Samuel W French, Linda Sher, Jeong Joseph Hyeongnam, and Hidekazu Tsukamoto. Cancer stem cells generated by alcohol, diabetes, and hepatitis c virus. *Journal of Gastroenterology and Hepatology*, 27:19–22, 2012.
[53] Jingsong Wang, Michael A Maldonado, et al. The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases. *Cellular and Molecular Immunology*, 3(4):255–261, 2006.

[54] Jing Zhao, Ting-Hong Yang, Yongxu Huang, and Petter Holme. Ranking candidate disease genes from gene expression and protein interaction: a katz-centrality based approach. *PLoS One*, 6(9):e24306, 2011.

[55] Jacek M Zurada, Aleksander Malinowski, and Ian Cloete. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS’94*, volume 6, pages 447–450. IEEE, 1994.