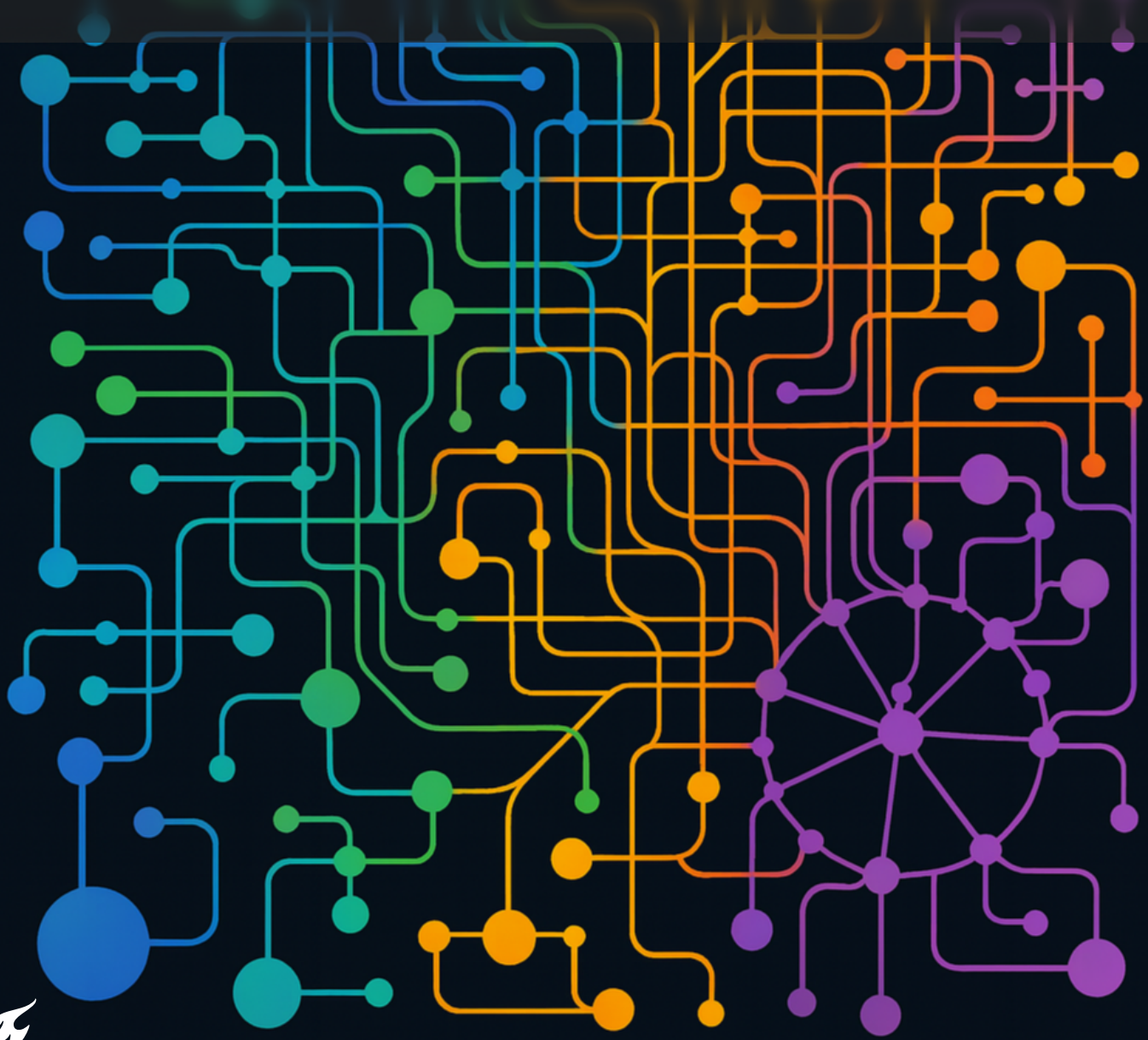
From Latent to Blatant Space

Coupling Biological Systems to Neural Networks for Improved Model Interpretability

MSc Thesis

Martijn Lieftinck



Delft University of Technology

TU Delft

From Latent to Blatant Space

Coupling Biological Systems to Neural Networks for Improved Model Interpretability

by

Martijn Lieftinck

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday July 16, 2025 at 13:00.

Student name:	Martijn Lieftinck
Student number:	4862856
Thesis committee:	
Prof. dr. ir. M.J.T. Reinders,	TU Delft, supervisor
T. Verlaan,	TU Delft, co-supervisor
Dr. M. Khosla,	TU Delft, External Committee Member
Faculty:	Electrical Engineering, Mathematics and Computer Science
Department:	Intelligent Systems
Research group:	Pattern Recognition and Bioinformatics
Lab:	Delft Bioinformatics Lab
Project duration:	November 2024 – July 2025

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Acknowledgments

I would like to express my gratitude to Prof. dr. ir. Marcel Reinders for giving me the opportunity to conduct this research. His supervision, critical view and constructive feedback were of great value. I especially appreciate the freedom he offered to choose a research direction that I found most interesting. I truly enjoyed doing this research under his supervision.

I also would like to thank Timo Verlaan, my daily supervisor, for his constant engagement. I enjoyed the many brainstorm sessions and exchanges of tips and tricks, which really helped me improve my work, as well as my academic qualities. I learned a lot during the many hours we spent in your office.

Furthermore, I would like to thank Dr. Megha Khosla for taking part in my Thesis Committee and evaluating my work.

Lastly, I would like to thank all my friends, family, colleagues, fellow students, and members of the Delft Bioinformatics Lab who showed genuine interest in my research while keeping me sharp by constantly challenging me to explain my work on every possible level of complexity. I genuinely enjoyed it.

Martijn Lieftinck
Delft, July 2025

From Latent to Blatant Space

Coupling Biological Systems to Neural Networks
for Improved Model Interpretability

M.A. Lieftinck July 7, 2025

Abstract

Deep Neural Networks (DNNs) are renowned for their high accuracy and versatility, which has led to their application in many fields of research, including biology. However, this accuracy often comes at the expense of interpretability, making it challenging to reason about the inner workings of most DNNs. Particularly in biological research, understanding the mechanisms behind specific outcomes is highly valuable. To elucidate the latent space of DNNs in the context of cancer biology, we introduce GONNECT: a Gene Ontology-derived Neural Network for Explainable Cancer Typing. GONNECT incorporates biological prior knowledge from the Gene Ontology (GO) directly into its network architecture, enabling interpretability through model structure. Using an autoencoder framework, we evaluate GONNECT as both encoder and decoder module and demonstrate its ability to learn which biological processes are distinctive for different cancer types. Furthermore, we show how a variant including soft links (GONNECT-SL) can expand on current knowledge by proposing new interactions between biological processes. GONNECT is flexible both in the amount of prior knowledge it incorporates and the set of input genes, and can potentially be applied in modeling of gene perturbation effects and drug target discovery.

Introduction

Deep learning has revolutionized machine learning research and applications. This field of machine learning has proven its versatility through the array of models that have emerged in recent years, from the first multilayer perceptron (MLP) models to state-of-the-art transformers [1]. Especially in the current AI era, deep learning is ubiquitous in research, business, education, and everyday life. Whereas deep neural networks (DNNs) are famous for their accuracy and flexibility, they are infamous for their lack of interpretability, as is often illustrated by their reference as “black box" models [2].

The inherent trade-off between accuracy and interpretability plays an important role in choosing the right model for a specific task [3], [4]. In biological predictive modeling, understanding the mechanisms underlying functional outcomes is as important as the predictions themselves. [5], [6]. Gaining insight into what the key components in a biological system are and how they interact to cause a specific phenotype is of great importance in advancing our understanding of biology and mechanisms of disease.

Many studies have attempted to improve the interpretability of deep learning models for biology by leveraging prior knowledge of biological systems. There are many variations of these biologically-informed neural networks (BINNs). Graph-based BINNs use gene interaction networks to structure genetic data and apply graph neural networks to exploit known gene interactions [7], [8].

ODE-based BINNs aim to model system behavior over time using known ordinary differential equations (ODEs). ODEs describing binding dynamics can be leveraged in knowledge-derived activation functions [9], [10]. Other ODE-based BINNs use ODEs to apply additional constraints in the loss function to ensure that the model output does not violate the known dynamics of the system [11]–[13].

Architecture-based BINNs use hierarchical ontologies directly as neural network architecture, coupling each network node to an ontology term. Links between nodes exist only if the associated ontology terms are also linked. Due to the hierarchical nature of these ontologies, they align relatively well with the layered structure of a multilayer perceptron (MLP).

Architecture-based BINNs create interpretable embeddings of the input features, where node activations reflect the contribution of the associated term to the predicted outcome or embedding. Coupling the inner workings of the model with known biological systems results in more transparent models with a smaller parameter space compared to their “black box" counterparts, allowing reasoning about predictions while reducing the amount of required training data.

Previous work has shown promising results for architecture-based BINNs in both performance and

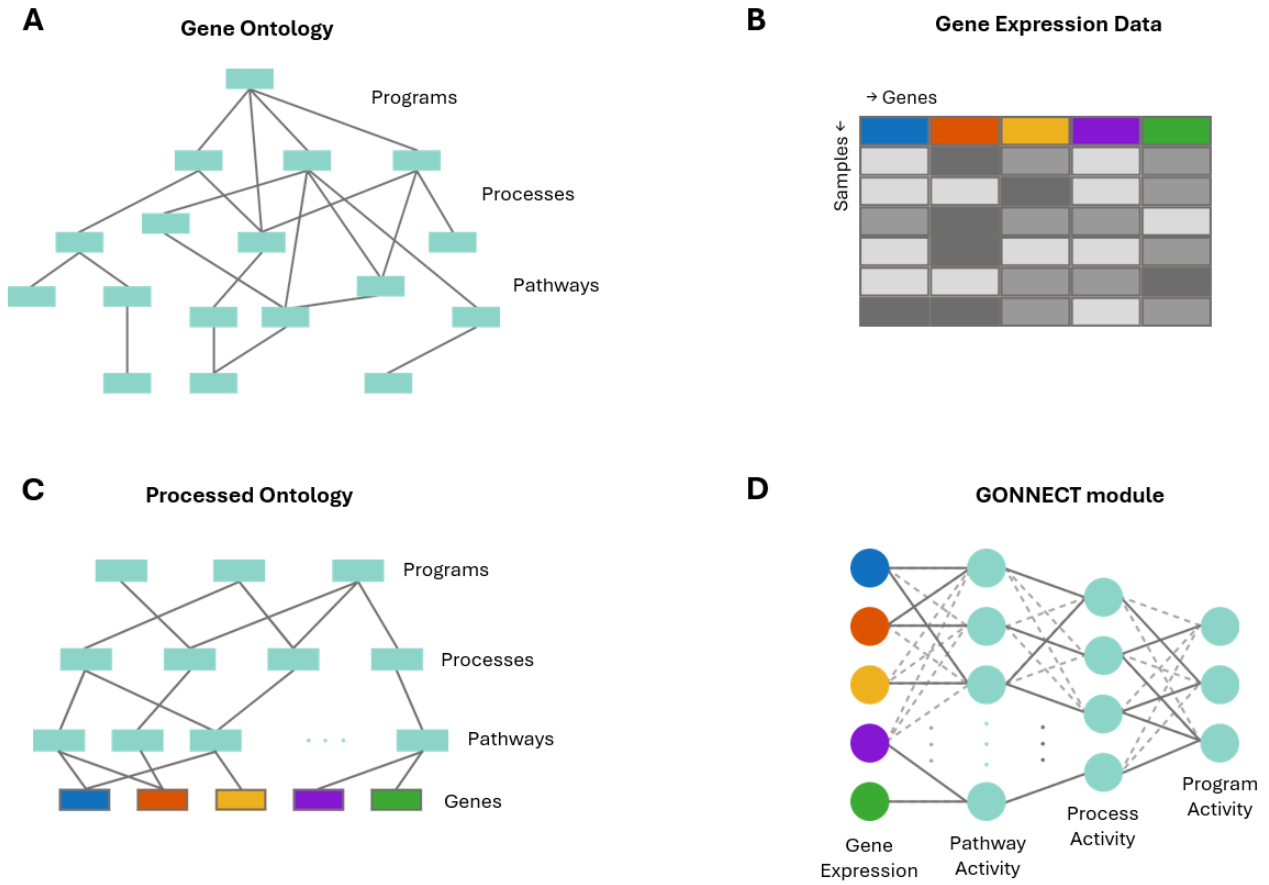


Figure 1: Overview of GONNECT design. **A**) Gene Ontology (GO) [14]: A hierarchical ontology graph containing experimentally verified prior knowledge on biological entities (GO-terms) and their relationships. **B**) Dataset containing gene expression data of different cancer types. **C**) Processed GO graph where genes from the dataset are linked to GO-terms based on associations between gene product and biological entity. **D**) GONNECT module where the network architecture is based on the processed GO graph. GONNECT takes gene expression as input and outputs values for the activity of GO-terms at the top of the hierarchy. Solid lines represent "fixed links": edges based on experimental evidence of a biological relationship. Dashed lines represent "soft links": unverified yet potential relationships between GO-terms.

interpretability [15]–[18]. These models are more efficient than their biologically-agnostic counterparts (i.e. standard MLPs), needing less parameters to obtain similar accuracy. However, a drawback of these prediction models is that they are trained to predict a specific phenotypic property. As a result, interpretation at the term level is only relevant in the context of that phenotype, precluding broader insight into general inter-term relationships.

To overcome the task-specificity of prediction-based approaches, autoencoder models aim to create a more generally interpretable embedding of the input features. Hierarchical prior knowledge can be used to embed quantitative gene expression data in terms of abstract ontology terms. The low-dimensional and interpretable embeddings can subsequently be used for different tasks such as classification, drug response prediction, and *in silico* modeling of gene perturbation effects [19]–[21]. Interestingly, all BINN autoencoder models use prior knowledge solely in the decoder module, meaning the embeddings that form the basis of interpretability are shaped by a "black box" encoder.

The concept of using experimentally verified prior knowledge to construct BINNs ensures a certain baseline understanding of the ground truth. However, it is also a limitation. The fields of biology from which this knowledge is obtained are still very active, meaning that information we assume to be the ground truth is incomplete and still subject to extension and revision. Especially in architecture-based models, the assumption of complete prior knowledge can lead to the exclusion of possibly important biological relationships, simply because they have not been verified yet. So far, there has been one study that addressed this problem: ExpiMap, a model that learns gene expression embeddings for single-cell reference mapping,

uses a biologically-informed single-layer decoder to reconstruct the embedded “gene programs" into the individual genes that make up those programs [19]. Since there might be genes involved in a certain gene program that are not included in the prior knowledge, ExpiMap is allowed to add new links between gene and gene program to improve performance.

Here we introduce GONNECT, a Gene Ontology-derived Neural Network for Explainable Cancer Typing. We use hierarchical prior knowledge from Gene Ontology (GO) [14], a literature-curated reference database, to construct an architecture-based BINN for use as encoder, decoder, or both in an autoencoder framework (Figure 1). Using self-supervised learning, we aim to obtain a multipurpose, biologically interpretable latent space where every node is associated with an ontology term (GO-term), thus representing a biological pathway, process, or program. Edges between nodes represent relationships between GO-terms. GONNECT is evaluated based on reconstruction performance of gene expression input, embedding quality, and latent space interpretation. To address the notion of incomplete prior knowledge, we developed a “soft link" variant, GONNECT-SL, where prior knowledge is augmented by allowing soft links: new relationships which are not present in GO.

Adaptive ontology processing enables training on any gene set with at least one GO-annotation. Furthermore, GONNECT can incorporate a variable number of GO-terms, determined by user-defined term-degree thresholds (i.e. limits on the number of parents/children per term). The flexibility to use GONNECT and GONNECT-SL as both encoder and/or decoder allows for a comparative analysis of the effects on performance and interpretability of introducing biological priors in the different autoencoder modules. Meanwhile, GONNECT-SL can learn new relationships through soft links not only between genes and ontology terms, but also between ontology terms themselves, providing suggestions for novel intra- and inter-process relationships.

Results

Gene Ontology and dataset

We use Gene Ontology (GO) [14] to supply GONNECT with hierarchical biological prior knowledge (general information on GO is available in Supplementary Information, Section S1). Before use in models, the GO ontology graph is filtered and processed (see Methods-Ontology processing).

The models in this study were trained and evaluated using publicly available data from The Cancer Genome Atlas (TCGA) [22]. TCGA contains gene expression data on human tumor samples. After preprocessing, the TCGA-derived dataset contained 9,797 samples, including 32 different cancer types, and their 1000 most highly variable genes with at least one GO-annotation (see Methods-Data preprocessing). The distribution of cancer types in the preprocessed dataset is available in Table S1.

GONNECT architecture

The processed GO graph determines the number of layers, nodes, and weights of GONNECT. GO processing produces a condensed graph suitable as neural network architecture (see Methods-Ontology processing). The processed graph comprises 623 GO-terms, 2,369 proxy terms, and 1,000 genes. The resulting GONNECT modules have an input dimension of 1,000, five network layers, and an embedding dimension of 109.

To evaluate the effects of incorporating biology into different autoencoder modules (encoder and/or decoder), GONNECT was tested in three configurations (Figure 2). One in which a GONNECT encoder is coupled to a biologically-agnostic, fully connected MLP decoder (Figure 2A), a second configuration consisting of an MLP encoder and GONNECT decoder (Figure 2B), and a third in which both the encoder and decoder are GONNECT modules (Figure 2C). In all three configurations, the embedding space is coupled to a GONNECT module, meaning that the embedding space is always associated with GO-terms.

We first developed GONNECT models where prior knowledge is a hard requirement for edge existence, meaning that there can only be an edge between nodes, and therefore a learnable weight, if the associated GO-terms are linked in the ontology graph (see Methods-Model implementations). These “fixed link" models are limited by the knowledge in GO. To address the fact that GO is an incomplete knowledge base that might not contain all relevant biological interactions, we also developed GONNECT-SL: a variant containing soft links (Figure 2A, dashed lines). GONNECT-SL models allow non-zero weights between terms that do not have a known, verified GO-relationship (see Methods-Model implementations).

The possibility of using additional edges enables augmentation of the prior knowledge graph with new, data-driven relationships. Soft link weights are regularized to favor known GO-edges, yet can activate new connections when they significantly reduce reconstruction loss (see Methods-Training).

GONNECT achieves better reconstruction performance than biologically-agnostic models

We compared the performance of GONNECT models to biologically-agnostic reference models. The first

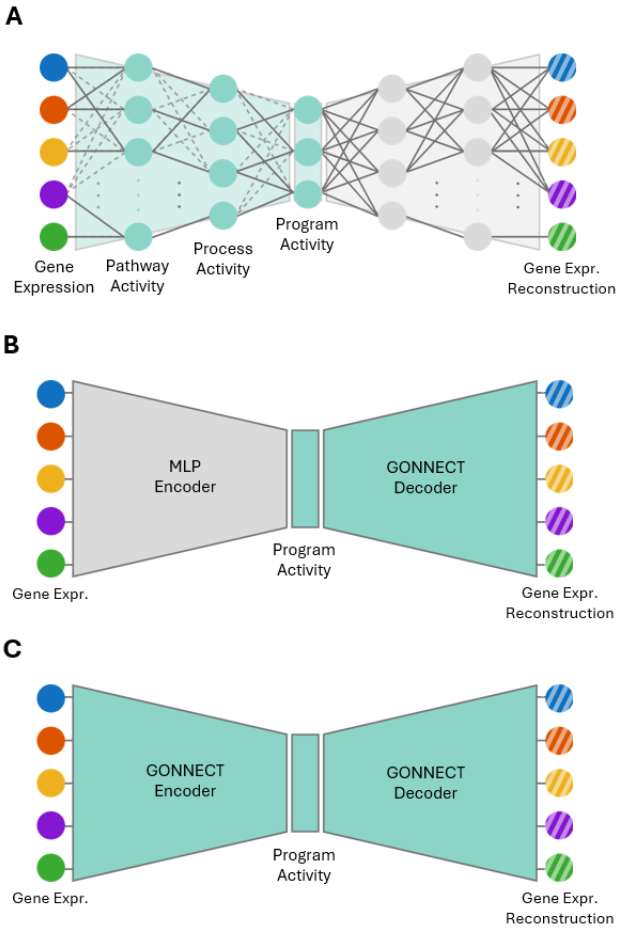


Figure 2: Schematic overview of different GONNECT configurations. **A**) Autoencoder with a GONNECT encoder such that network nodes represent GO-terms and solid links represent known relationships between terms. The decoder is a fully connected MLP with an identical but mirrored node layout. **B**) Configuration where the encoder is an MLP and the decoder is a GONNECT module. **C**) Configuration where both encoder and decoder are GONNECT modules, such that every node in the network is coupled to a GO-term.

reference model is an MLP with the same number of layers and nodes per layer as GONNECT, but fully connected. The result is a network with the same number of weight matrices as GONNECT, the same weight matrix sizes, but the amount of learnable weights is two orders of magnitude higher ($3 \cdot 10^6$ opposed to $1 \cdot 10^4$).

The second reference model, GONNECT-R, is similar to regular GONNECT, except that the links between nodes are randomized in a degree-preserving manner (see Methods-Model implementations). The resulting model has the same number of parameters and topological properties as GONNECT, but lacks the biological foundation.

All models were trained to minimize the mean square error (MSE) between input and reconstruction (see Methods-Training). For each GONNECT variant, we evaluated the average reconstruction loss over five model instances (Figure 3A). GONNECT-SL achieved the lowest MSE, outperforming the other GONNECT variants as well as fully connected MLPs. Regular GONNECT performed better compared to GONNECT-R in all autoencoder configurations, indicating that biological context contributes to performance.

The MSE per cancer type showed that some cancer types consistently showed a higher MSE than others, especially in fixed link models (Figure 4). However, this pattern is probably not related to the use of biological prior knowledge, since it was equally present in the biologically-agnostic GONNECT-R models.

Autoencoder configuration influences reconstruction performance

For each GONNECT variant (GONNECT, GONNECT-SL and GONNECT-R), we tested three different autoencoder configurations (Figure 2). The GONNECT encoder in fixed link models performed significantly better than the GONNECT decoder in fixed link models, indicating that strictly using GO-derived edges in the decoder negatively impacts autoencoder performance (Figure 3A). Remarkably, across all three variants, the GONNECT encoder consistently performed on par with the fully connected MLP, despite using fewer parameters. Both GONNECT-SL encoders and decoders benefit from the addition of soft links, as both outperform their fixed link counterparts. However, the performance improvement is significantly larger for the decoder than for the encoder, closing the performance gap between configurations when soft links are allowed.

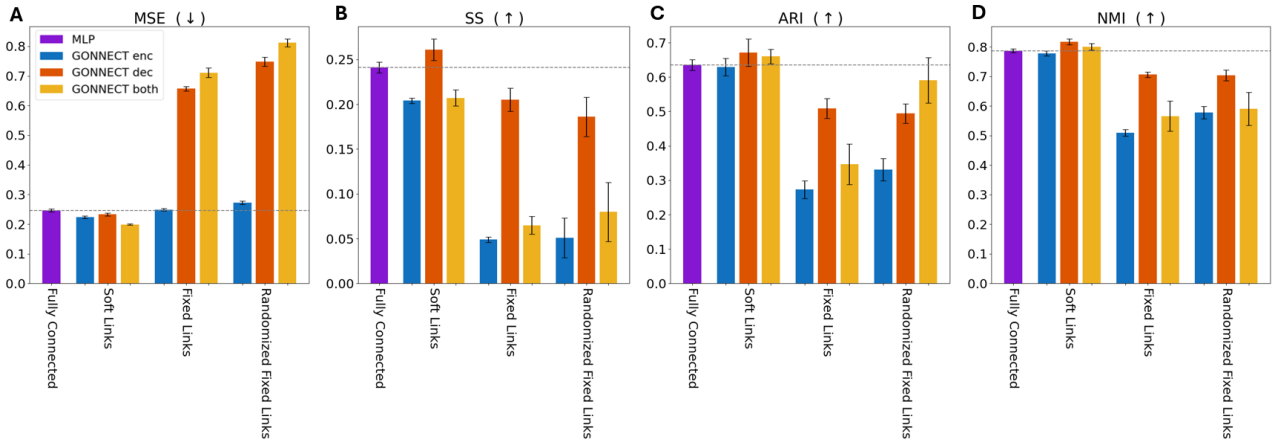


Figure 3: Performance metrics for different GONNECT variants. Colors indicate model configuration, where 'enc' is short for encoder, 'dec' for decoder and 'both' for encoder and decoder. Bar height denotes the mean over five independently trained model instances, error bars denote the standard deviation, and arrows indicate whether the metric improves by ascending or descending in value. Variants are grouped by the type of links between network nodes. The horizontal dashed line indicates the MLP reference **A**) Mean square error (MSE) of gene expression reconstruction. **B**) Silhouette score (SS) of the embedding space. **C**) Adjusted rand index (ARI) of the embedding space. **D**) Normalized mutual information (NMI) of the embedding space (see Methods-Embedding Metrics).

GONNECT largely preserves cancer type separability in the embedding space

With the aim of obtaining an interpretable model that can be used in multiple biologically relevant prediction tasks, we evaluated the quality of the learned embedding space using three clustering metrics: Silhouette score (Figure 3B), adjusted rand index (Figure 3C) and normalized mutual information (Figure 3D) (see Methods-Embedding Metrics).

The soft link models outperformed both fixed link variants in all metrics, with the GONNECT-SL decoder even surpassing the fully connected MLP. Although GONNECT-R appears to perform slightly better than GONNECT, its high variance suggests a less consistent embedding space organization.

The GONNECT decoders scored higher than GONNECT encoders, although the effect for soft link models is smaller than for fixed link models. However, when a GONNECT decoder is coupled to a GONNECT encoder, the embedding scores drop again. The effect of the encoder module seems stronger than that of the decoder module, since the performance of dual GONNECT models is closer to that of GONNECT encoders than of decoders.

We also qualitatively compared the embedding spaces of different GONNECT configurations using a two-dimensional t-SNE transform [23] of the embedded dataset, as well as the original input space (Figure 5). Additional visualizations using different dimensionality reduction methods are available in Figures S2 and S3.

Clusters in the high-dimensional input space are mostly conserved in GONNECT embeddings, enabling the separation of most cancer types. Overlapping clusters are cancer types from similar tissues (ESCA, STAD, COAD and READ all occur in the digestive system; UCEC, UCS, CESC and OV all occur in the female reproductive system), resulting in a relatively similar gene expression profile for these cancer types, which explains why they appear in the same neighborhood of the embedding space. Cancer type information is available in Table S1.

GONNECT encoder models (Figure 5B and 5D) produce a more homogeneously scattered t-SNE compared to MLP encoders (Figure 5A and 5C), indicating reduced cluster separability. Meanwhile, the type of decoder has a minimal effect on the organization of the embedding space.

Combining results from the clustering metrics and the t-SNEs shows that GONNECT decoders only slightly lower embedding quality compared to MLPs, whereas GONNECT encoders significantly impair embedding quality. Combining GONNECT encoder and decoder offers a small improvement over encoder-only models.

Node activations identify associated cancer types

In GONNECT, biological knowledge constrains network connectivity, enforcing interpretable and biologically relevant predictions. Through training, the model learns the activity of these GO-derived edges,

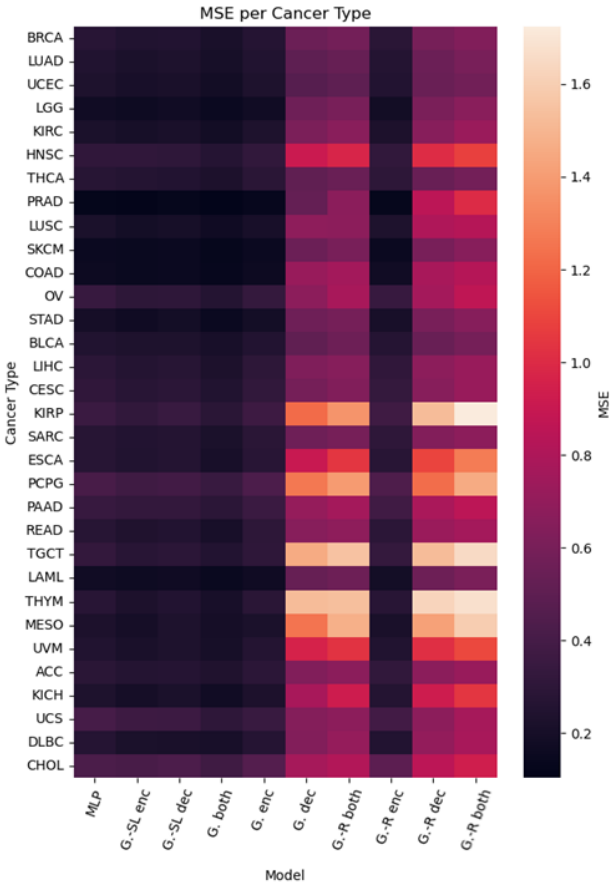


Figure 4: Reconstruction error per cancer type of different GONNECT variants. Error is expressed as mean square error (MSE). Cancer types are ordered by descending number of samples in the dataset. GONNECT is abbreviated with 'G.', encoder with 'enc' and decoder with 'dec'.

reflecting the importance of each of the biological relationships between ontology terms. In turn, the importance of these relationships determines which nodes become active for a given input. By analyzing the activations of latent nodes during the forward pass of trained models, we can get insight into which GO-terms are most active for each cancer type.

Activation polarity does not model mode of action

GONNECT showed similar node activation magnitudes per cancer type for different model instances. This indicates that GONNECT repeatedly learned to consider the same GO-terms as active for a given cancer type. However, the polarity of the activations differed per model instance, caused by random weight initialization and no constraints on the sign of weights during training. As a result, the signs of activations cannot be used as an indication of the mode of action of the associated interaction, which means that negative activations indicate neither inhibitory nor activating interactions. Activation data is available in Figures S4 and S5.

GONNECT encoder captures both tissue and cancer biology

The interpretability of node activations was evaluated using ROC-AUC scores of a selection of terms that are expected to vary in activity over the different cancer types in the dataset (see Methods-Latent node activation analysis). Information on cancer types and selected GO-terms is available in Tables S1 andS2 respectively.

The GONNECT encoder was able to discriminate cancer types that are known to have aberrant activity of a certain biological process using the activation of the network node associated with that process (Figure 6A). Most of these highly predictive terms were related to general tissue biology: Cholesterol metabolism is predominantly active in liver tissue and adrenal glands for the production of bile acid and steroids [24], [25]. This is reflected in the ROC-AUC score of the GO-term *cholesterol metabolic process* for LIHC and ACC, and *bile acid metabolic process* for LIHC (Figure 6A, asterisks). LIHC is also well distinguished from other types by the GO-term *negative regulation of blood coagulation*, which is known to be influenced by the liver [26]. Other examples of congruence between the ROC-AUC score and tissue biology are found for neuron-associated terms: *positive regulation of myelination* and *positive regulation of neuron projection regeneration* are especially predictive of LGG samples.

In addition to tissue-specific biology, the GONNECT encoder also appeared to have captured cancer biology information: BLCA is best predicted by *C21-steroid hormone metabolic process* (Figure 6A, circles). We

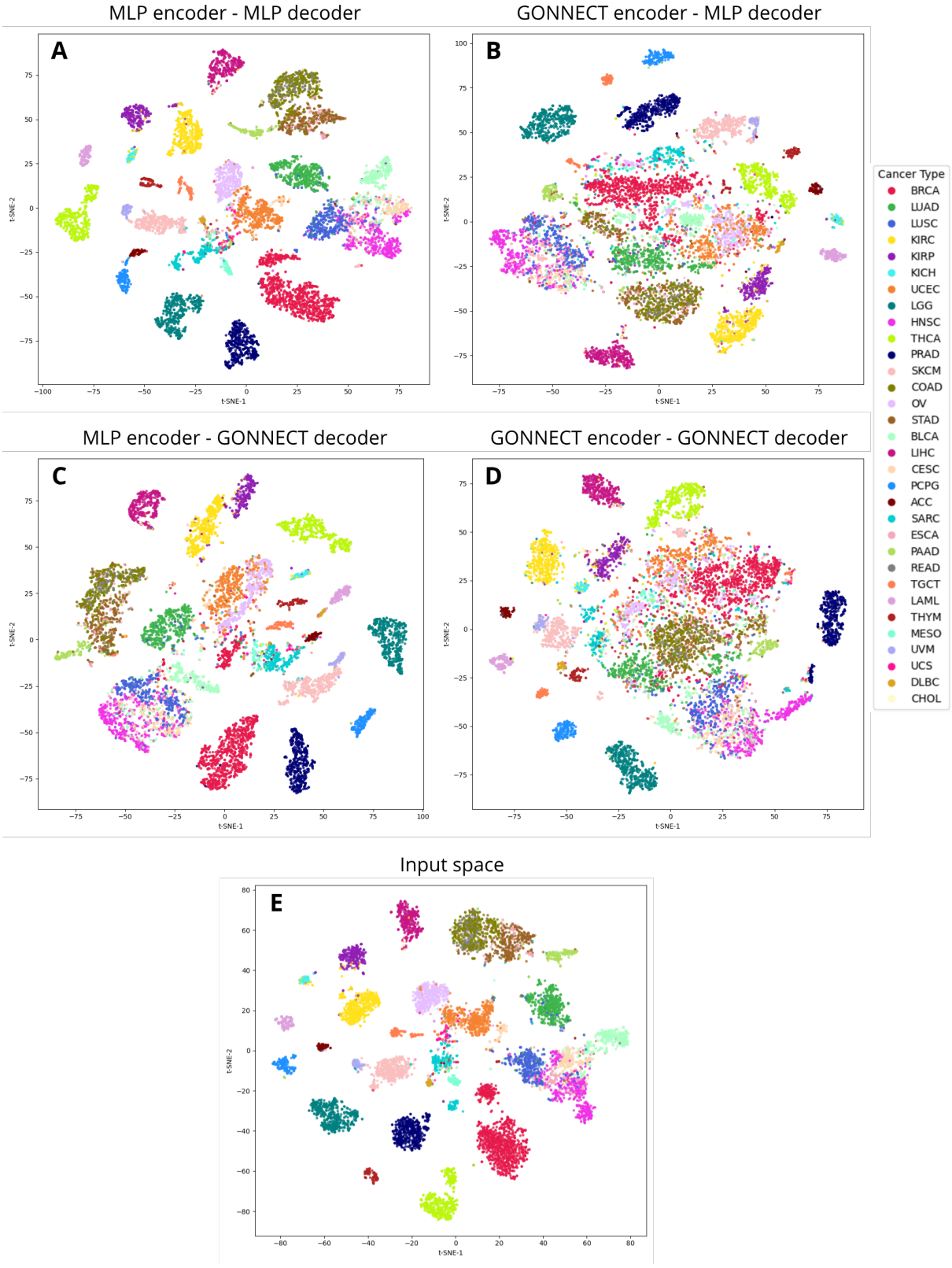


Figure 5: Two-dimensional t-SNE transform [23] of the embedding space learned by different GONNECT configurations. Samples are labeled by cancer type. **A**) Embedding space of the fully connected model where both encoder and decoder are MLPs. **B**) Embedding space of a GONNECT encoder with MLP decoder. **C**) Embedding space of an MLP encoder with GONNECT decoder. **D**) Embedding space of a GONNECT encoder with GONNECT decoder. **E**) The t-SNE transform of the original high-dimensional input space.

could not find evidence in literature of aberrant steroid metabolism in healthy bladder tissue. However, a recent study indicates a depletion of steroids in male bladder tumors [27]. Similarly, *positive regulation of MAP kinase activity* is activated in UVM cancers [28]. We also looked at terms that separate different cancer types from the same tissue. For lung samples, we found *DNA damage checkpoint signaling* to be the term with the largest score difference between LUAD and LUSC, which is also one of the processes influenced by the set of differentially expressed genes between LUAD and LUSC [29].

Not all ROC-AUC scores could be supported by current literature. Some terms are known to be upregulated in multiple cancer types, such as *extracellular matrix organization* for PAAD and SARC [30], while only one of them obtained a significant ROC-AUC score for that GO-term (Figure 6A, triangles). ROC-AUC scores of all 623 GO-terms per GONNECT module are available in Figures S6 and S7.

GONNECT decoder struggles to reconstruct specific processes

The GONNECT decoder shows a different distribution of ROC-AUC scores compared to the GONNECT encoder. If we compare the same examples, we only see agreements for LIHC and LGG (Figure 6B). Furthermore, many GO-terms show similar patterns in the ROC-AUC scores per cancer type. On further inspection, similar score patterns belong to similar GO-terms that are variants of the same type of biological process. The GONNECT decoder appears to struggle to differentiate in the activations of these related terms, given that it must reconstruct them from one common ancestor term.

Soft links propose novel biological interactions

The ontology used in this study as a source of prior knowledge is not complete. GO is constantly revised and expanded with new information gathered from experiments. It is therefore a rather bold assumption to use GO as ground truth for the existence of relationships between biological entities. Many relationships might exist, but have yet to be experimentally verified, and are therefore not present in GO.

GONNECT-SL balances performance and prior knowledge dependence

The introduction of soft links allows for data-driven augmentation of incorporated prior knowledge. The process of learning soft links must be carefully controlled to prevent too many new edges from being incorporated, which would degrade the biological interpretability in favor of reconstruction. The soft link models were optimized to have similar performance as fully connected models using a minimal amount of active soft links (see Methods-Soft link tuning). The results of soft link hyperparameter tuning are available in Figures S8 and S9.

The weights of the resulting GONNECT-SL model can be divided into three groups: GO-derived links, active soft links, and inactive soft links. GO-derived links are not subject to regularization and have a similar distribution to fixed link weights (Figure 7). Of the available soft links, 96% have a magnitude $|w| < 0.001$, and only 174 (0.005%) soft links have a weight magnitude $|w| > 0.01$ and are therefore considered active. This relatively small amount of active soft links was sufficient to perform on par with fully connected models.

Active soft links are supported by literature

To interpret GONNECT-SL, we gathered the most active soft links in each hierarchical layer in the network and looked for potential biological relevance. Interestingly, the ten most active soft links between genes and GO-terms (the lowest hierarchical layer) were all linked to the endoplasmic-reticulum-associated protein degradation pathway (ERAD). According to GO, ERAD links to only one gene in the dataset. However, some of the genes linked through soft links are likely to be related to ERAD as well: Q9BZQ8 encodes a protein involved in the integrated stress response, a pathway that is activated by accumulation of unfolded proteins in the endoplasmic reticulum (ER) [31], [32]. Q6UXG2 is also involved in the unfolded protein response in the ER [33]. Q9BPY8 encodes co-chaperone proteins that assist in refolding and/or degradation of proteins during cellular stress [34]. However, other genes that obtained a high soft link weight were less likely to be involved in ERAD, e.g., genes that play a role in cholesterol transport, insulin sensitivity, or fibril formation.

Soft links between GO terms in higher hierarchical layers also provided valuable insight into possibly unknown interactions. In addition to linking genes involved in integrated stress response to the ERAD GO-term, GONNECT-SL also linked ERAD to the *integrated stress response* GO-term in the consecutive layer. There are multiple studies that point to the influence of ER stress and unfolded protein response on the ERAD pathway [31]–[33], [35]. Another interesting soft link that became active during training is that between *positive regulation of epithelial to mesenchymal transition* (EMT) and *tissue development*. In literature, it is well described how EMT plays a crucial role in embryonic tissue development [36]. In fact, these two GO-terms are known to be related, but with a relation-

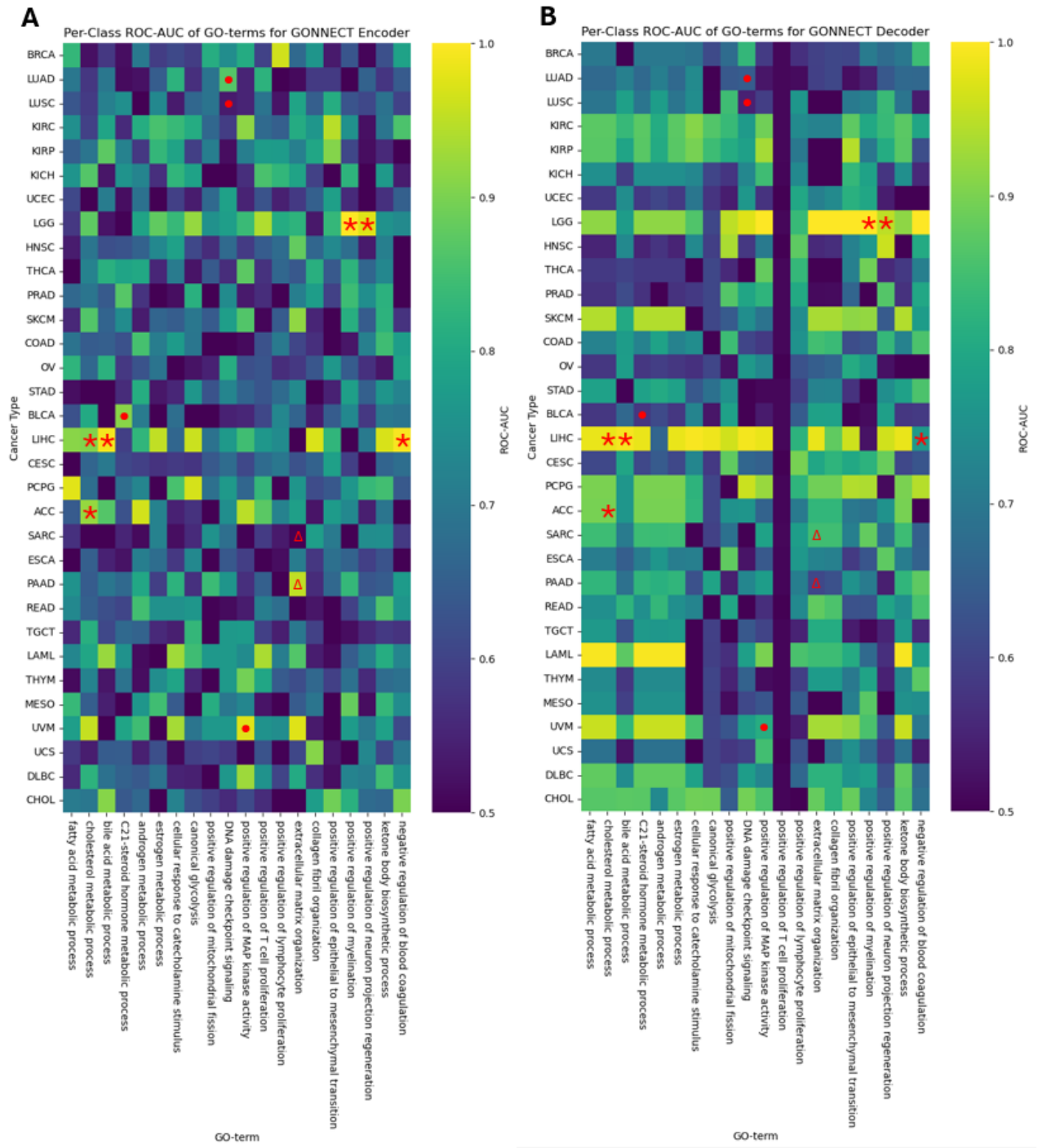


Figure 6: Heatmap of ROC-AUC scores of linear regression models trained on the activations of GONNECT nodes to distinguish cancer types (see Methods-Latent node activation analysis). Red asterisks indicate processes related to general tissue biology, red circles indicate processes related to cancer biology and red triangles indicate where similar scores were expected based on cancer type. **A**) ROC-AUC scores from GONNECT encoder nodes. **B**) ROC-AUC scores from GONNECT decoder nodes.

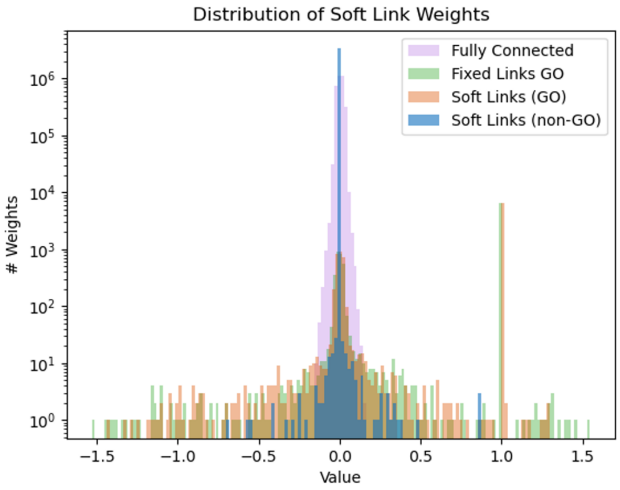


Figure 7: Distribution of learned weights of GONNECT-SL. For comparison, the weight distribution of a fully connected MLP (purple), and the distribution of regular GONNECT (green) is shown. The weights of GONNECT-SL are split up into the weights of links that were already present in GO (orange) and those that are not, meaning they can become active soft links (blue). The spike at 1.0 represents unlearnable weights, fixed a a value of 1.0 (see Methods-Model implementations).

ship that was removed during ontology processing (see Methods-Ontology processing). GONNECT-SL thus learned a connection it had never seen before but in reality has been proven to exist.

Interestingly, most active soft links contain terms that are involved in cancer, such as those related to ER stress [37], [38], EMT [39], and angiogenesis [40], as well as more general terms such as DNA repair and adaptive immune response.

Active soft links form hubs on low-degree nodes

In both the encoder and the decoder, one node participates in more than half of the ten most active soft links between genes and GO-terms. This node thus forms a hub for active soft links, having active outgoing soft links as well. Interestingly, these hub nodes originally had very few incoming edges (one for the encoder hub, four for the decoder hub), which means that they carried relatively little information about the input before the addition of soft links.

GONNECT trains less efficient than MLPs

Despite having orders-of-magnitude fewer parameters, GONNECT required more time to train relative to standard MLPs. The GO-derived constraints give GONNECT a sparse architecture with less learnable weights. However, the average time per epoch increased for both GONNECT and GONNECT-SL models from 1.1 s for MLPs, to an average of 7.2 s for GONNECT variants.

The number of epochs until convergence remained similar between MLPs and GONNECT encoder models, but increased four times for decoder models and 17 times for dual GONNECT configurations. Furthermore, additional memory was required to store the GO-derived masks used during training, resulting in an additional 200MB of memory used (see Methods-Training). Training statistics for the different models are available in Table S3.

Discussion

Deep neural networks would greatly benefit from methods that can interpret their internal representations. In biology, the ability to trace back the mechanisms involved in a particular prediction is crucial for research on the mechanisms of disease and medicine. Here we introduced GONNECT, a sparse, biologically-informed neural network that leverages prior knowledge from Gene Ontology to provide interpretability of the latent space of gene expression autoencoders, in terms of biological processes and their relationships.

Biological prior knowledge improves performance and explainability

By mapping each neuron to a GO-term and wiring edges to match GO relationships, GONNECT produces a directly interpretable network. On gene expression reconstruction, GONNECT outperforms similar models with an equal amount of parameters, but without a biologically-informed architecture. Our results show that GONNECT can recognize up- or down-regulated pathways and processes based on gene expression data and use these activities to distinguish cancer types. Its soft link extension, GONNECT-SL, surpasses the biologically-agnostic reference models, including those with significantly more parameters, while the learned soft links provide relevant suggestions for novel biological relationships between genes and ontology terms.

GONNECT performs best when used as encoder

GONNECTs built-in flexibility to be used as both encoder and decoder allowed us to compare different autoencoder configurations in terms of performance and explainability of the inner workings of the model. So far, all existing biologically-informed autoencoders use prior knowledge exclusively in the decoder. Our results show that GONNECT performs best as encoder module. The GONNECT encoder achieved better

reconstruction and a more interpretable latent space than decoders. Although GONNECT decoders had a more organized embedding space, they were not able to fully exploit the embeddings, resulting in a higher reconstruction loss.

The fact that GONNECT encoders can achieve better reconstruction from a lower-quality embedding space shows that the MLP decoder in these models is more powerful than a GONNECT decoder. Similarly, the high-quality embeddings of GONNECT decoder models show how MLP encoders produce better embeddings than GONNECT encoders. Interestingly, the effect on reconstruction seems stronger for decoders than for encoders. It could therefore be the case that the high reconstruction performance of GONNECT encoder models is explained by being coupled to a more potent MLP decoder.

Node activations are partially explainable

One way to interpret GONNECTs inner workings is to conceptualize it as simulating the internal processes of a cell. The activation of each neuron then reflects the activity of its corresponding process, and network weights indicate the strength of influence between processes. In GONNECT models, the signs of the activations proved uninformative, as different model instances showed consistent magnitudes of activation but arbitrary signs. We showed how some processes exhibited distinctive magnitudes of activation for samples in which we know that these processes have a distinct activity compared to other cell types. However, we also observed up- or down-regulated processes in cancer types that are not known to be associated with that process, as well as unobtrusive activations for processes that were expected to distinguish a cancer type.

One explanation for these observations is that GONNECT learned cancer type characteristics that have not yet been discovered or described in literature. An unexpectedly high ROC-AUC score for a process-cancer type pair could reveal novel associations, while a lower than expected score could indicate that other cancer types share similar process activity, but the behavior might only have been described for one cancer type.

Alternatively, during optimization, GONNECT may focus on a subset of genes and processes to help distinguish cancer types, while downplaying those emphasized in literature. Unrelated or less studied genes and processes might show a similar expression pattern between cancer types as known markers, resulting in redundancy in the sets of genes and processes that can be used to distinguish a certain cancer type.

An important limitation in dealing with this redundancy is the lack of validation data. There is no ground truth available on how active a certain process is, given the expression of its associated genes. Therefore, we cannot guide GONNECT to prefer biologically plausible processes when expression patterns correlate, nor can we confirm that it has learned correct relations between gene expression and process activity.

Soft links might be biased by topology and dataset

GONNECT-SL showed a consistent tendency to concentrate active soft links on a single GO-term. The resulting hub node combines many different activations and passes the aggregate to many others. It would be interesting to study whether this behavior is biologically justifiable or if it is general network behavior for this approach to soft link learning.

Moreover, the predominance of cancer-related terms among active soft links suggests that training on cancer expression data introduced a bias in GONNECT toward expanding cancer biology interactions. Although GO is not cancer-oriented, the model may have become biased to augment interactions related to cancer biology, as a result of using TCGA data. This would suggest that GONNECT learned what biological processes play an important role in cancer.

Implementation of sparsity hindered learning efficiency

Despite its sparse connectivity, GONNECT required more time and more memory than fully connected models. Transforming GO to a structure that is suitable as MLP required extensive processing and resulted in artifacts that hindered training: in order to define distinct network layers, proxy terms were added to the ontology graph. As a result, 60% of the nodes in the final graph were proxy terms. Since these proxy terms do not have biological meaning, they cannot participate in optimization. This was achieved by automatically resetting the weight towards all proxy terms to one and their bias to zero, after each optimization step. This intervened with optimization, partially disrupting gradient flow during backpropagation. As a result, both time per epoch and number of epochs were negatively affected.

Additionally, the implementation of GONNECT required more memory compared to fully connected models, since the GO-derived masks used during training had to be stored as additional model attributes. An implementation using sparse COO tensors for both weight matrices and masks was considered but rejected due to a doubling in runtime complexity.

Outlook and applications

We applied GONNECT and GONNECT-SL as autoencoder modules, aiming to train a task-invariant model with biological knowledge and general interpretability. We used cancer-oriented data to train and evaluate our models and obtained results that were related

to cancer as well. It would be interesting to study how GONNECT behaves after training on a different dataset with healthy samples: Do cancer-associated terms lose their predictive power? Do the same soft links become active? To what extent were the results of this study biased by the dataset used?

In terms of implementation, improvements in weight masking could lead to more efficient optimization, potentially increasing accuracy. Furthermore, an approach in which the GONNECT decoder has more freedom to distinguish child terms from a single parent could improve decoder interpretability.

GONNECT could prove a valuable asset in the fields of computational and experimental biology. A model that can accurately simulate the activity of, and interactions between the many complex and abstract biological processes in a cell, solely from gene expression data, would greatly benefit the field of mutational research and drug development. GONNECT lends itself well to modeling multiple gene knockouts and overexpression of genes. In turn, this could be used to assess potential drug targets and predict drug side effects. Furthermore, if the genes involved in a certain process are not yet fully identified, disruption of that process can still be simulated by manipulating the corresponding node activation.

GONNECT can also be applied in combination with a classification head, directly using the learned embeddings, or by retraining in a more task-specific setting. GONNECT-SL can help to find new relationships, not only between genes and pathways, but also between more abstract processes. Highly active soft links could give direction for experimental research into unknown biological dependencies.

The lack of need for labels means that GONNECT can be easily retrained with different datasets, opening up the possibility to train personalized models that could be used in personalized medicine.

GONNECT and GONNECT-SL have shown promising results in generating biologically substantiated outcomes, providing both performance and interpretability, while giving suggestions for knowledge gaps. A further optimized approach could improve performance, interpretability, and efficiency. The opportunities for improvement in combination with a wide potential applicability foreshadow a lasting relevancy of GONNECT and its successors in advancing the field of biologically-informed neural networks.

Methods

Ontology processing

The biological prior knowledge used as model architecture was derived from Gene Ontology[1] (GO) [14]. First, GO-terms were filtered by namespace to select just the Biological Process terms. The relationships between ontology terms were then filtered for *is_a* relationships, to ensure that the resulting ontology satisfied the properties of a directed acyclic graph (DAG).

The ontology graph was extended with the genes from the dataset using the gene annotation file[2] (GAF) for *Homo sapiens*. GO-terms without any genes linked to their subtree were removed from the DAG (Figure 8A).

Increasing average connectivity

Terms with low connectivity were removed to reduce the size of the ontology while maintaining its biological interpretability. Based on user-defined thresholds, any term with fewer parent terms than the *parent threshold* and fewer children than the *child threshold*, were merged into their parent(s). This merge operation links all children of the merged term to all parents of the merged term and subsequently removes the term itself from the ontology (Figure 8B). Since the merge operation alters parent-child relationships, multiple rounds are needed to cover all cases where a term only met the merge requirements as a consequence of a previous merge operation (e.g. a child of a term was merged in a previous round, resulting in the number of remaining children of that term to fall below the threshold, in turn resulting in the term in question being merged as well). This process of iterating over all terms and testing for the merge conditions was repeated until convergence.

An unwanted byproduct of the merge operation is the formation of skip connections, where a single edge connects a member from a term its subtree to a member of said term its supertree, allowing a bypass of the term itself. These skip connections were removed, or pruned, after each round of merges, where a round is one loop over all terms. The merge-prune operations condense the ontology graph, resulting in a more compact, yet still interpretable form.

Next to the merge conditions for minimal number of parents and children, we also included a third condition to control the number of depth levels in the graph and, therefore, the number of network layers in the resulting GONNECT module. This *depth population*

[1] `https://purl.obolibrary.org/obo/go/go-basic.obo` Date of access: 27 Oct 2024

[2] `https://current.geneontology.org/annotations/goa_human.gaf.gz` Date of access: 10 Sep 2024

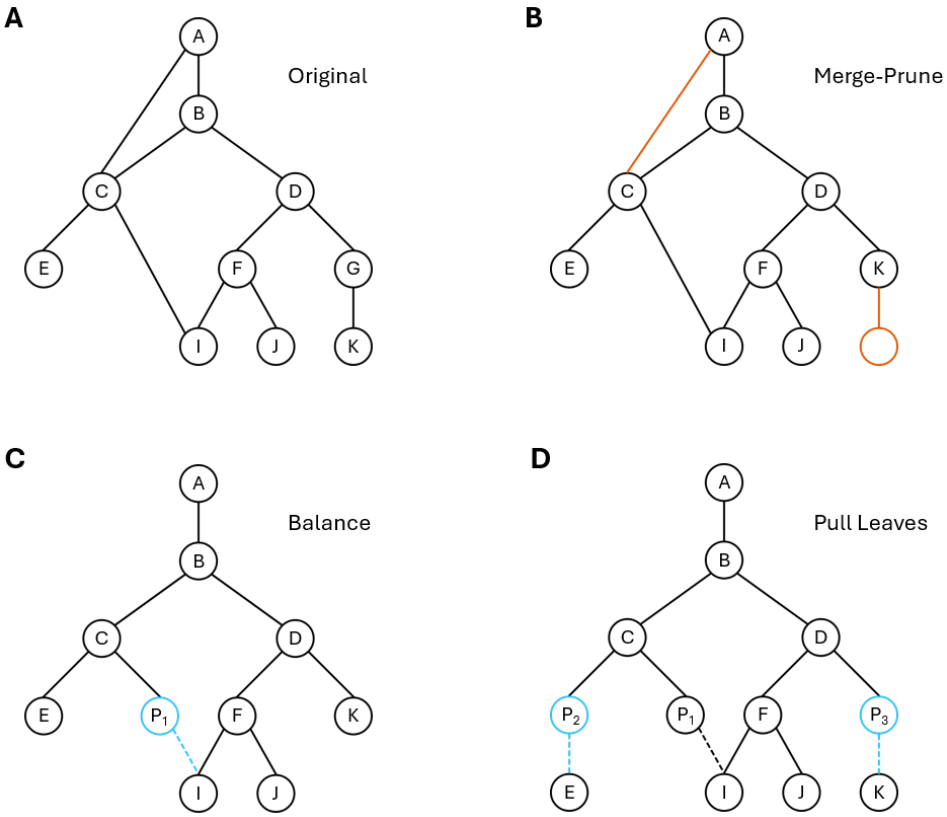


Figure 8: Example of the different GO processing steps. Nodes and edges that are colored red in a panel are removed from the graph during that processing step, those colored blue are added to the graph. Dotted lines represent edges towards proxy terms. **A**) An graph with similar properties as GO. Nodes E, I, J and K represent genes that have been linked to their annotated GO-terms. **B**) Effect of the merge-prune operation for a parent threshold of 1 and a child threshold of 2. Node G gets merged into node D and the skip connection AC gets removed. **C**) The balancing operation introduces a proxy term P_1 to balance node I. **D**) Gene nodes E and K are pulled down by introducing proxy terms P_2 and P_3, such that all genes are at maximum depth.

threshold determines the minimum number of GO-terms at any depth level. If the number of terms is below the threshold, all terms at that depth level are merged into their parents. The graph used in this study was processed with the following merge conditions:

$$\begin{aligned} \text{parent threshold} &= 1 \\ \text{child threshold} &= 30 \\ \text{depth population threshold} &= 50 \end{aligned}$$

Neural network compatibility

For a graph to be used as the architecture of an MLP, we identified the following graph properties as requirements:

- Acyclic: All possible paths through the graph should start at a root node and end in a leaf node, or vice versa, and have no loops (i.e. a network pass passes each layer once)
- Balanced: All possible paths from one node to another should have an equal amount of edges (i.e. a term occurs in only one single network layer)
- Equal path lengths from root to leaves: All leaf nodes should have an equal path length towards the root layer (i.e., all input nodes are present in the first network layer)

Acyclicity is ensured by using only *is_a* relationships of GO and, therefore, does not need to be considered explicitly.

To balance the ontology graph, the graph was traversed depth-first and a proxy term (a new term inserted between the original term and a subset of its children) was inserted wherever the current term had imbalanced children, of which the current term was on the shortest branch (Figure 8C). This process was repeated for multiple traversals until the graph converged.

Equal path lengths from root to leaf were achieved

by inserting proxy terms above genes until the depth level of the gene was equal to the maximum depth of the graph (Figure 8D). The resulting graph has a layout that can be used directly as an MLP, without the need for residual connections.

Data preprocessing

The dataset used for training and evaluation was constructed and downloaded from the Genomic Data Commons (GDC) portal [41]. Samples were included if they were open access, part of the TCGA program, and had available gene expression quantification data. The retrieved dataset contained 10,498 samples from 9,648 patients, including transcripts per million (TPM) data on 59,427 genes. Next, each TPM value v_i of gene i was log-transformed as follows:

$$v_i \to \log(v_i + 1)$$

UniProt IDs were retrieved for each gene name using UniProts online ID mapping tool[3] [42]. Genes without a match were dropped, after which the gene names of the remaining genes were replaced with their Uniprot ID. The resulting IDs were intersected with the set of Uniprot IDs that have at least one GO-annotation in the Biological Process namespace according to the gene association file (GAF). Samples from healthy tissue and genes with zero variance were dropped, leaving 9797 samples containing 17,491 genes. In light of computational costs, data availability, and model complexity, we trained our models on the 1,000 most highly variable genes in the dataset, selected using the Seurat procedure [43].

Model implementations

Each term in the ontology is represented by a node in the neural network, and each relationship between two terms is modeled with a learnable weight between the two corresponding nodes. All models are implemented in Python using pytorch (v 2.5.1). Parsing of the OBO file[4] containing the ontology was performed using the goatools package (v 1.4.12) [44].

GONNECT module

The biologically-informed architecture is implemented through sparse weight matrices, where a nonzero element represents an edge between nodes, and thus a relationship between their associated GO-terms. These weight matrices are obtained by applying GO-derived masks to the weights of each network layer.

[3] https://www.uniprot.org/id-mapping

[4] https://current.geneontology.org/ontology/go-basic.obo

The terms in the processed ontology graph were grouped by depth, after which adjacency matrices between consecutive depth levels were constructed based on parent-child relationships. These adjacency matrices were in turn used to mask the corresponding weight matrices of GONNECT. In addition to these edge masks $\mathbf{M_e}$, we also store a 1D proxy mask $\mathbf{M_b}$ per GONNECT layer. The proxy mask is used to fixate weights towards proxy terms to 1, and the bias of proxy terms to 0. This ensures unaffected signal transduction through proxy terms and preserves interpretability. The resulting forward pass is defined as

$$\mathbf{x}^{(i+1)} = \text{ReLU}\left(\mathbf{W_M}^{(i)}\,\mathbf{x}^{(i)} + \mathbf{b_M}^{(i)}\right)$$

where i denotes the network layer, $\mathbf{x}$ the input and $\mathbf{W_M}$ and $\mathbf{b_M}$ are the masked versions of the weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$, for which the following holds

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if} \quad \mathbf{M_b}_i = 1 \\ \mathbf{W}_{ij} & \text{else} \end{cases} \tag{1}$$

$$\mathbf{W_M} = \mathbf{W} \odot \mathbf{M_e} \tag{2}$$

$$\mathbf{b_M} = \mathbf{b} \odot \mathbf{M_b} \tag{3}$$

where $\odot$ denotes the Hadamard product. At model initialization, the weight matrices of GONNECT are filled by drawing from a Kaiming uniform distribution [45]. After initialization and after each optimization step during training, weight masking is repeated to ensure that the masked values remain fixed.

GONNECT-SL module

Like regular GONNECT, GONNECT-SL stores similar edge masks and proxy masks. After model initialization using a Kaiming normal distribution, these masks are used to fix the weights toward proxy terms to 1 (Eq. 2), and proxy biases to 0 (Eq. 3). However, instead of setting non-GO weights to 0, they are initialized as soft link from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.001$. Weights and biases of proxy terms are kept constant by reapplying masks after each optimization step during training.

GONNECT-R module

One of the baseline models used to evaluate performance is GONNECT-R: a variant of GONNECT that preserves the number of nodes, layers, and nodes per layer, as well as node connectivity. GONNECT-R shuffles the GO-derived edges between nodes while preserving in- and out-degree. The resulting model loses its biological interpretability but largely retains the architectural properties of the GONNECT module from which it originates.

Edge shuffling was performed using a Monte Carlo algorithm that picks two directed edges and swaps their sink nodes, similar to XSwap [46]. Between each pair of network layers, a total of QE swaps were performed, where E is the number of GO-derived edges between the two layers, and $Q = 100$, a heuristic to assume randomness [47].

Training

The autoencoder models are trained using the mean squared error (MSE) of input reconstruction, given by

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$$

Where $\mathbf{x}_i$ is the original input sample, $\hat{\mathbf{x}}_i$ is the reconstructed output of the autoencoder, n is the number of samples, and $\|\cdot\|_2^2$ denotes the squared ℓ_2-norm.

For soft link models, an additional term is added to the MSE that penalizes the absolute value of all soft links (i.e. weights representing edges that are not present in the ontology graph). The resulting loss function $\mathcal{L}_{SL}$ is used to train these models.

$$\mathcal{L}_{\text{SL}} = \mathcal{L}_{\text{MSE}} + \alpha \left(\frac{1}{m}\sum_{j=1}^{k} \left\| \mathbf{W}^{(j)} \odot (1 - \mathbf{M_e}^{(j)}) \right\|_1 \right)$$

Where α is the regularization factor (see Methods-Soft link tuning), $\mathbf{W}^{(j)}$ the weight matrix corresponding to the j-th network layer, $\mathbf{M_e}^{(j)}$ the edge mask for the j-th network layer, $\odot$ denotes the Hadamard product, k the number of network layers, m the total number of soft links in the network and $\|\cdot\|_1$ denotes the ℓ_1-norm. The motivation for weight regularization using the ℓ_1-norm is to allow some soft links to receive a relatively high weight while keeping the majority of soft link weights close to zero.

All models were trained to minimize their respective loss functions through stochastic gradient descent (SGD). Using random train-validation-test splits of 70%-15%-15%, each model was trained multiple times with different split initializations, using early stopping with patience parameter $p = 10$.

Soft link tuning

Control over the number of participating soft links is achieved with hyperparameter α. This α weights the ℓ_1-norm of the soft link weights against the reconstruction loss, which means that the higher α, the larger the contribution of soft link magnitudes to the overall training loss, and thus the stronger the regularization effect. In extremes, a soft link model where $\alpha = 0$ is equivalent to a fully connected MLP, while $\alpha = \infty$ is equivalent to a fixed link model. The α hyperparameter was optimized to maximize reconstruction performance while minimizing the number of non-zero soft link weights. A value of $\alpha = 1 \cdot 10^3$ resulted in a model in which 174 non-GO links were active as soft links, meaning that they obtained a weight w for which $|w| > 0.01$ after training for 1000 epochs.

Embedding Metrics

The quality of the embeddings is evaluated based on three metrics. The silhouette score (SS) uses intra- and inter-cluster distances to indicate how tight and well-defined a certain clustering is [48]. The ground truth cancer type labels were used to define the clusters. The SS can take values between -1 and 1, where SS ≈ 1 indicates tightly and well separated clusters, SS ≈ 0 indicates cluster overlap, and SS ≈ -1 indicates wrong cluster assignments.

The adjusted rand index (ARI) measures the pairwise agreement between two clusterings [49]. The clusterings used for comparison are the ground truth labels for cancer type and an unsupervised k-means clustering of the embedding space, where k is equal to the number of cancer types in the dataset. If the two clusterings are identical, ARI $= 1$. For a random clustering, ARI ≈ 0.

Normalized mutual information (NMI) gives the shared information between different clusterings [50]. Again, the ground truth cancer type labels were compared with a k-means clustering where k was equal to the number of cancer types. A score of NMI $= 0$ indicates no mutual information, and NMI $= 1$ indicates perfect correlation between the two clusterings.

Latent node activation analysis

The interpretability of node activations is evaluated by their ability to distinguish different cancer types. For each node in GONNECT that is coupled to a GO-term, we trained a linear regression model on the mean activation of that node per cancer type in a one-vs-rest setup. Each node received a score equal to the area under the ROC curve (ROC-AUC) obtained by evaluating the linear regression models on the same test set as used in model training. This ROC-AUC score was used to express the ability of a node to distinguish a certain cancer type from other cancer types.

Resource Availability

All code used in data processing, model construction and training, and analysis is available at `https://github.com/mlieftinck/Thesis_BINN/`. The data used in this study is open access and available at `https://portal.gdc.cancer.gov/`.

References

[1] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021. DOI: 10.1007/s42979-021-00592-x.

[2] W. R. Ashby, "An introduction to cybernetics," pp. 86–117, 1956.

[3] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 248–266, ISBN: 9781450393522. DOI: 10.1145/3531146.3533090.

[4] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: A comparative study of machine learning models," *Journal of Reliable Intelligent Environments*, vol. 11, no. 1, p. 1, 2025. DOI: 10.1007/s40860-024-00240-0.

[5] V. Chen, M. Yang, W. Cui, J. S. Kim, A. Talwalkar, and J. Ma, "Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments," *Nature methods*, vol. 21, no. 8, pp. 1454–1461, 2024. DOI: 10.1038/s41592-024-02359-7.

[6] D. Sidak, J. Schwarzerová, W. Weckwerth, and S. Waldherr, "Interpretable machine learning methods for predictions in systems biology from omics data," *Frontiers in Molecular Biosciences*, vol. 9, 2022. DOI: 10.3389/fmolb.2022.926623.

[7] C. Liu, A. H. Wan, H. Liang, *et al.*, "Biological informed graph neural network for tumor mutation burden prediction and immunotherapy-related pathway analysis in gastric cancer," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 4540–4551, Jan. 2023, ISSN: 20010370. DOI: 10.1016/j.csbj.2023.09.021.

[8] R. K. Tripathy, Z. Frohock, H. Wang, *et al.*, "An explainable graph neural network approach for integrating multi-omics data with prior knowledge to identify biomarkers from interacting biological domains," 2024. DOI: 10.1101/2024.08.23.609465.

[9] A. Nilsson, J. M. Peters, N. Meimetis, B. Bryson, and D. A. Lauffenburger, "Artificial neural networks enable genome-scale simulations of intracellular signaling," *Nature Communications*, vol. 13, 1 Dec. 2022, ISSN: 20411723. DOI: 10.1038/s41467-022-30684-y.

[10] I. Hossain, V. Fanfani, J. Fischer, J. Quackenbush, and R. Burkholz, "Biologically informed neuralodes for genome-wide regulatory dynamics," *Genome Biology*, vol. 25, 1 Dec. 2024, ISSN: 1474760X. DOI: 10.1186/s13059-024-03264-0.

[11] J. H. Lagergren, J. T. Nardini, R. E. Baker, M. J. Simpson, and K. B. Flores, "Biologically-informed neural networks guide mechanistic modeling from sparse experimental data," *PLoS Computational Biology*, vol. 16, 11 Dec. 2020, ISSN: 15537358. DOI: 10.1371/journal.pcbi.1008462.

[12] G. Massonis, A. F. Villaverde, and J. R. Banga, "Distilling identifiable and interpretable dynamic models from biological data," *PLoS Computational Biology*, vol. 19, 10 October Oct. 2023, ISSN: 15537358. DOI: 10.1371/journal.pcbi.1011014.

[13] M. Sadria and V. Swaroop, "Discovering governing equations of biological systems through representation learning and sparse model discovery," 2024. DOI: 10.1101/2024.09.19.613953.

[14] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000. DOI: 10.1038/75556.

[15] J. Ma, M. K. Yu, S. Fong, *et al.*, "Using deep learning to model the hierarchical structure and function of a cell," *Nature Methods*, vol. 15, pp. 290–298, 4 Apr. 2018, ISSN: 15487105. DOI: 10.1038/nmeth.4627.

[16] J. Hao, Y. Kim, T. K. Kim, and M. Kang, "Pasnet: Pathway-associated sparse deep neural network for prognosis prediction from high-throughput data," *BMC Bioinformatics*, vol. 19, 1 Dec. 2018, ISSN: 14712105. DOI: 10.1186/s12859-018-2500-z.

[17] H. A. Elmarakeby, J. Hwang, R. Arafeh, *et al.*, "Biologically informed deep neural network for prostate cancer discovery," *Nature*, vol. 598, pp. 348–352, 7880 Oct. 2021, ISSN: 14764687. DOI: 10.1038/s41586-021-03922-4.

[18] Y. Hao, J. D. Romano, and J. H. Moore, "Knowledge-guided deep learning models of drug toxicity improve interpretation," *Patterns*, 2022, ISSN: 26663899. DOI: 10.1016/j.patter.2022.100565.

[19] M. Lotfollahi, S. Rybakov, K. Hrovatin, *et al.*, "Biologically informed deep learning to query gene programs in single-cell atlases," *Nature Cell Biology*, vol. 25, pp. 337–350, 2 Feb. 2023, ISSN: 14764679. DOI: 10.1038/s41556-022-01072-x.

[20] B. M. Kuenzi, J. Park, S. H. Fong, *et al.*, "Predicting drug response and synergy using a deep learning model of human cancer cells," *Cancer Cell*, vol. 38, 672–684.e6, 5 Nov. 2020, ISSN: 18783686. DOI: 10.1016/j.ccell.2020.09.014.

[21] D. Doncevic and C. Herrmann, "Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations," *Bioinformatics*, vol. 39, 6 Jun. 2023, ISSN: 13674811. DOI: 10.1093/bioinformatics/btad387.

[22] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (tcga): An immeasurable source of knowledge," *Contemporary Oncology/Współczesna Onkologia*, pp. 68–77, 2015, ISSN: 1428-2526. DOI: 10.5114/wo.2014.47136.

[23] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html.

[24] C. L. Cummins, D. H. Volle, Y. Zhang, *et al.*, "Liver x receptors regulate adrenal cholesterol balance," *The Journal of Clinical Investigation*, vol. 116, no. 7, pp. 1902–1912, Jul. 2006. DOI: 10.1172/JCI28400.

[25] A. D. Petrescu, J. Kain, V. Liere, T. Heavener, and S. DeMorrow, "Hypothalamus-pituitary-adrenal dysfunction in cholestatic liver disease," *Frontiers in endocrinology*, vol. 9, p. 660, 2018. DOI: 10.3389/fendo.2018.00660.

[26] V. E. Leonardi Filippo Maria Nicola De, "Anticoagulation in cirrhosis: A new paradigm?" *Clin Mol Hepatol*, vol. 23, no. 1, pp. 13–21, 2017. DOI: 10.3350/cmh.2016.0110.

[27] K. Kettunen, J. Mathlin, T. Lamminen, *et al.*, "Profiling steroid hormone landscape of bladder cancer reveals depletion of intratumoural androgens to castration levels: A cross-sectional study," *Ebiomedicine*, vol. 108, 2024. DOI: 10.1016/j.ebiom.2024.105359.

[28] S. E. Dickinson, E. R. Olson, J. Zhang, *et al.*, "P38 map kinase plays a functional role in uvb-induced mouse skin carcinogenesis," *Molecular carcinogenesis*, vol. 50, no. 6, pp. 469–478, 2011. DOI: 10.1002/mc.20734.

[29] D. Anusewicz, M. Orzechowska, and A. K. Bednarek, "Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of notch, hedgehog, wnt, and erbb signalling," *Scientific reports*, vol. 10, no. 1, p. 21 128, 2020. DOI: 10.1038/s41598-020-77284-8.

[30] M. Rafaeva, A. R. Jensen, E. R. Horton, *et al.*, "Fibroblast-derived matrix models desmoplastic properties and forms a prognostic signature in cancer progression," *Frontiers in Immunology*, vol. 14, p. 1 154 528, 2023. DOI: 10.3389/fimmu.2023.1154528.

[31] I. Novoa, Y. Zhang, H. Zeng, R. Jungreis, H. P. Harding, and D. Ron, "Stress-induced gene expression requires programmed recovery from translational repression," *The EMBO journal*, 2003. DOI: 10.1093/emboj/cdg112.

[32] K. Pakos-Zebrucka, I. Koryga, K. Mnich, M. Ljujic, A. Samali, and A. M. Gorman, "The integrated stress response," *EMBO reports*, vol. 17, no. 10, pp. 1374–1395, 2016. DOI: 10.15252/embr.201642195.

[33] M. Schröder and R. J. Kaufman, "The mammalian unfolded protein response," *Annu. Rev. Biochem.*, vol. 74, no. 1, pp. 739–789, 2005. DOI: 10.1146/annurev.biochem.73.011303.074134.

[34] J. H. Seo, J.-H. Park, E. J. Lee, *et al.*, "Ard1-mediated hsp70 acetylation balances stress-induced protein refolding and degradation," *Nature communications*, vol. 7, no. 1, p. 12 882, 2016. DOI: 10.1038/ncomms12882.

[35] C. Hetz, "The unfolded protein response: Controlling cell fate decisions under er stress and beyond," *Nature reviews Molecular cell biology*, vol. 13, no. 2, pp. 89–102, 2012. DOI: 10.1038/nrm3270.

[36] J. P. Thiery, H. Acloque, R. Y. Huang, and M. A. Nieto, "Epithelial-mesenchymal transitions in development and disease," *cell*, vol. 139, no. 5, pp. 871–890, 2009. DOI: 10.1016/j.cell.2009.11.007.

[37] Y. C. Tsai and A. M. Weissman, "The unfolded protein response, degradation from the endoplasmic reticulum, and cancer," *Genes & Cancer*, vol. 1, no. 7, pp. 764–778, 2010. DOI: 10.1177/1947601910383011.

[38] H. Kim, A. Bhattacharya, and L. Qi, "Endoplasmic reticulum quality control in cancer:
Friend or foe," in *Seminars in cancer biology*, Elsevier, vol. 33, 2015, pp. 25–33. DOI: 10.1016/j.semcancer.2015.02.003.

[39] Y. Nakaya and G. Sheng, "Emt in developmental morphogenesis," *Cancer letters*, vol. 341, no. 1, pp. 9–15, 2013. DOI: 10.1016/j.canlet.2013.02.037.

[40] T. Tonini, F. Rossi, and P. P. Claudio, "Molecular basis of angiogenesis and cancer," *Oncogene*, vol. 22, no. 42, pp. 6549–6556, 2003. DOI: 10.1038/35025220.

[41] M. A. Jensen, V. Ferretti, R. L. Grossman, and L. M. Staudt, "The nci genomic data commons as an engine for precision medicine," *Blood, The Journal of the American Society of Hematology*, vol. 130, no. 4, pp. 453–459, 2017. DOI: 10.1182/blood-2017-03-735654.

[42] T. U. Consortium, "Uniprot: The universal protein knowledgebase in 2025," *Nucleic Acids Research*, vol. 53, no. D1, pp. D609–D617, Nov. 2024, ISSN: 1362-4962. DOI: 10.1093/nar/gkae1010.

[43] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology*, vol. 33, no. 5, pp. 495–502, 2015. DOI: 10.1038/nbt.3192.

[44] D. V. Klopfenstein, L. Zhang, B. S. Pedersen, *et al.*, "Goatools: A python library for gene ontology analyses," *Scientific reports*, vol. 8, no. 1, p. 10 872, 2018. DOI: 10.1038/s41598-018-28948-z.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015. DOI: 10.48550/arXiv.1502.01852.

[46] S. Hanhijärvi, G. C. Garriga, and K. Puolamäki, "Randomization techniques for graphs," in *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)*, pp. 780–791. DOI: 10.1137/1.9781611972795.67.

[47] M. Espinoza, "On network randomization methods: A negative control study," *Fairfield, CT: Fairfield University*, 2012.

[48] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

[49] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985. DOI: 10.1007/BF01908075.

[50] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New journal of physics*, vol. 11, no. 3, p. 033 015, 2009. DOI: 10.48550/arXiv.0802.1218.

Supplementary Information

S1 Gene Ontology

The Gene Ontology (GO) knowledgebase is a comprehensive resource for computational analysis of molecular biology and genetics [1]. The hierarchical ontology is structured as a directed graph comprising terms that may have multiple parent and child terms. These GO-terms denote well-defined biological concepts and have a unique alphanumerical identifier, as well as a textual definition, making GO both human-readable and machine-readable. The many different GO-terms are connected through relationships such as *is_a* and *part_of*.

GO-terms are divided into three non-overlapping namespaces, each describing a different aspect of gene product biology:

- Molecular Function (MF): includes terms related to the activity of gene product at the molecular level, e.g. *catalysis* and *DNA binding*
- Cellular Component (CC): includes terms that denote where in the cell the gene product is active, e.g. *ribosome* and *cytoskeleton*
- Biological Process (BP): includes terms that describe larger processes that are the result of many molecular activities, e.g. *cell cycle* and *signal transduction*

GO makes use of annotation files to link gene products to GO-terms based on experimental evidence, sequence similarity, or phylogenetic relations. Annotations include one of the different types of relationship, a reference to the source of the annotation, and an evidence code, and they are manually curated by experts. Figure S1[1] depicts part of the BP namespace graph.

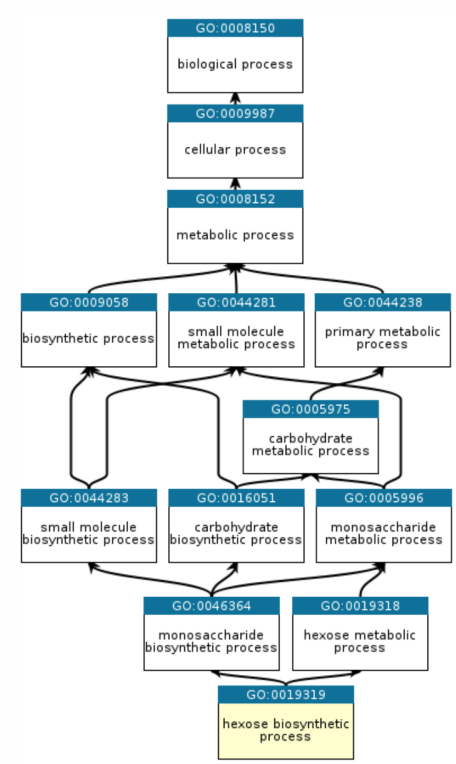


Figure S1: Section of the Gene Ontology graph. Depicted is the supertree of GO:0019319 (*hexose biosynthetic process*), which includes all parent terms based on *is_a* relationships up to the root term of the BP namespace: *biological process*.

[1]Figure from `https://geneontology.org/docs/ontology-documentation/`

Cancer type	Full name	Samples in dataset		Occurs in
BRCA	Breast invasive carcinoma	1060	10.8%	Breast
LUAD	Lung adenocarcinoma	516	5.3%	Lung
UCEC	Uterine corpus endometrial carcinoma	513	5.2%	Endometrium
LGG	Brain lower grade glioma	511	5.2%	Central nervous system
KIRC	Kidney renal clear cell carcinoma	508	5.2%	Kidney
HNSC	Head and neck squamous cell carcinoma	498	5.1%	Head and neck
THCA	Thyroid carcinoma	497	5.1%	Thyroid
PRAD	Prostate adenocarcinoma	477	4.9%	Prostate
LUSC	Lung squamous cell carcinoma	460	4.7%	Lung
SKCM	Skin cutaneous melanoma	440	4.5%	Skin
COAD	Colon adenocarcinoma	433	4.4%	Colon
OV	Ovarian serous cystadenocarcinoma	403	4.1%	Ovary
STAD	Stomach adenocarcinoma	397	4.1%	Stomach
BLCA	Bladder urothelial carcinoma	387	4.0%	Bladder
LIHC	Liver hepatocellular carcinoma	362	3.7%	Liver
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	293	3.0%	Cervix
KIRP	Kidney renal papillary cell carcinoma	282	2.9%	Kidney
SARC	Sarcoma	250	2.6%	Soft Tissue
ESCA	Esophageal carcinoma	180	1.8%	Esophagus
PCPG	Pheochromocytoma and paraganglioma	177	1.8%	Head and neck
PAAD	Pancreatic adenocarcinoma	173	1.8%	Pancreas
READ	Rectum adenocarcinoma	148	1.5%	Rectum
TGCT	Testicular germ cell tumors	138	1.4%	Testes
LAML	Acute myeloid leukemia	137	1.4%	Bone marrow
THYM	Thymoma	119	1.2%	Thymus
MESO	Mesothelioma	87	0.9%	Mesothelium
UVM	Uveal melanoma	77	0.8%	Eye
ACC	Adrenocortical carcinoma	76	0.8%	Adrenal glands
KICH	Kidney chromophobe	65	0.7%	Kidney
UCS	Uterine carcinosarcoma	53	0.5%	Uterus
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	45	0.5%	Lymphatic System
CHOL	Cholangiocarcinoma	35	0.4%	Bile Duct

Table S1: The different cancer types in the preprocessed TCGA [2] dataset. The table includes full name, distribution of samples in the dataset, and the location where each cancer type occurs in the human body.

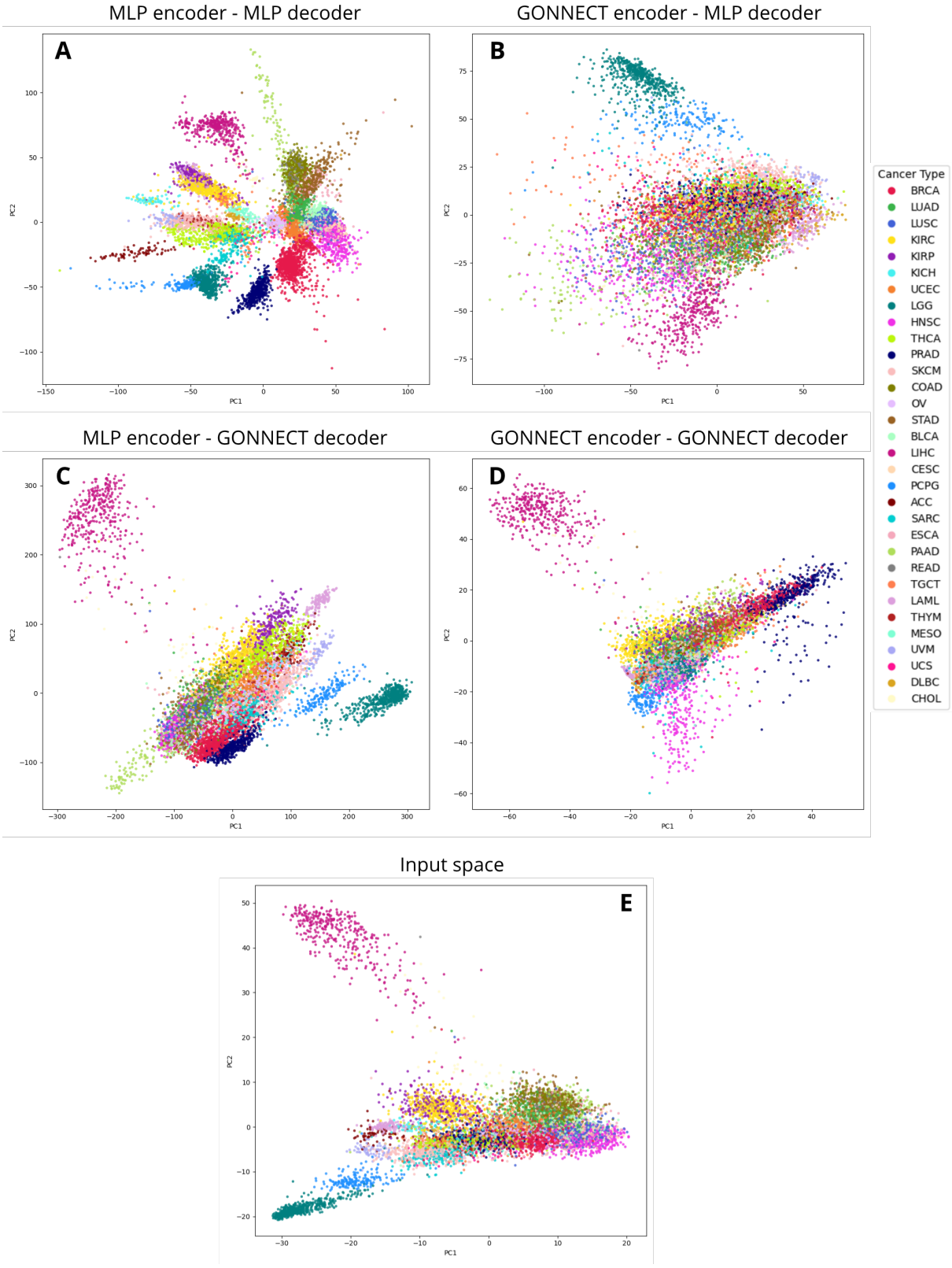


Figure S2: Principle component analysis (PCA) [3] of the embedding space learned by different GONNECT configurations. Samples are labeled by cancer type. **A**) Embedding space of the fully connected model where both encoder and decoder are MLPs. **B**) Embedding space of a GONNECT encoder with MLP decoder. **C**) Embedding space of an MLP encoder with GONNECT decoder. **D**) Embedding space of a GONNECT encoder with GONNECT decoder. **E**) PCA plot of the original high-dimensional input space.

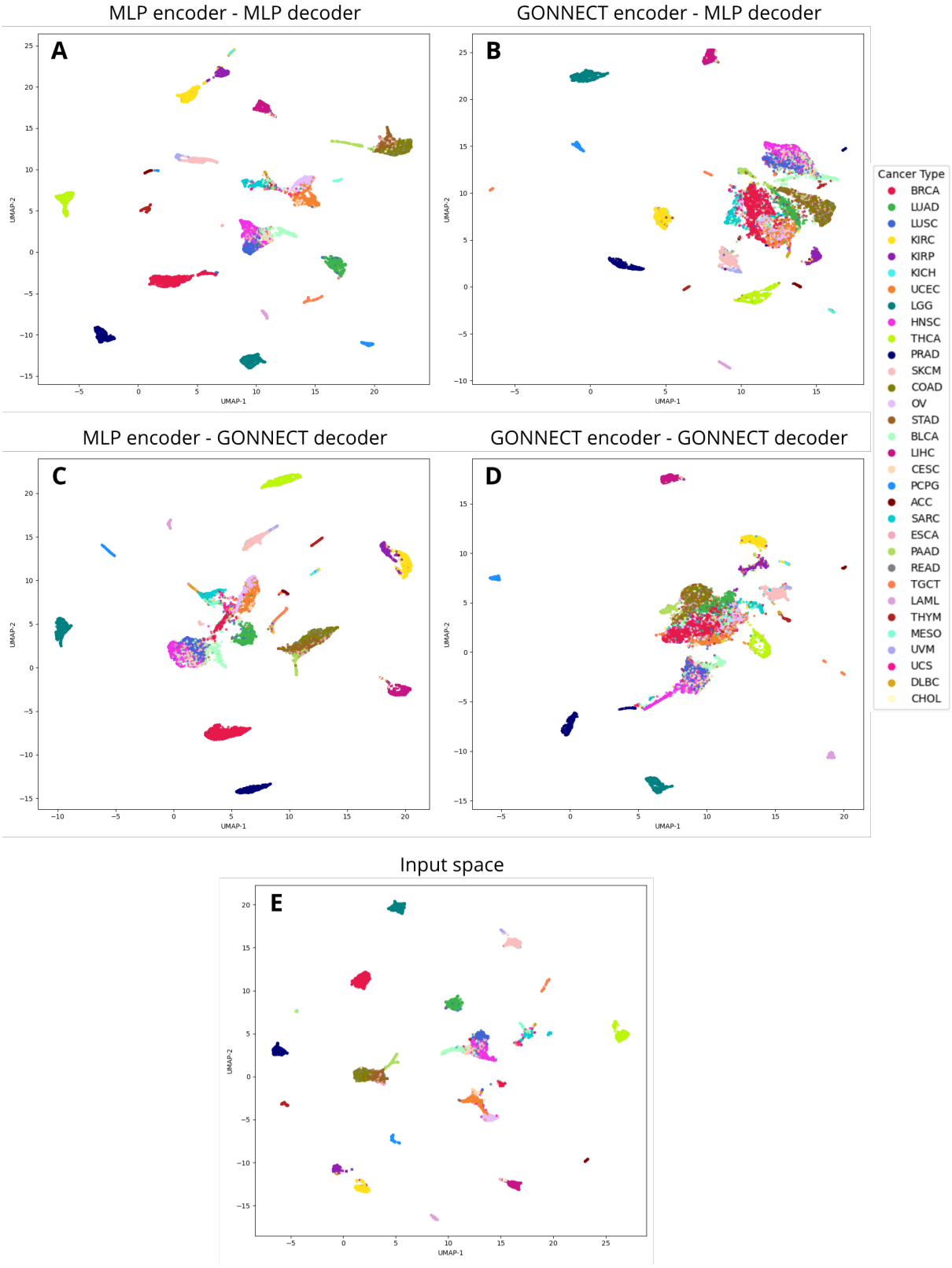


Figure S3: Two-dimensional UMAP transform [4] of the embedding space learned by different GONNECT configurations. Samples are labeled by cancer type. **A**) Embedding space of the fully connected model where both encoder and decoder are MLPs. **B**) Embedding space of a GONNECT encoder with MLP decoder. **C**) Embedding space of an MLP encoder with GONNECT decoder. **D**) Embedding space of a GONNECT encoder with GONNECT decoder. **E**) The UMAP of the original high-dimensional input space.

GO ID	Term name	Raised activity expected
GO:0006631	Fatty acid metabolic process	BRCA, LIHC, CHOL, OV, PCPG
GO:0008203	Cholesterol metabolic process	LIHC, CHOL, ACC
GO:0008206	Bile acid metabolic process	LIHC, CHOL
GO:0008207	C21-steroid hormone metabolic process	ACC, PCPG
GO:0008209	Androgen metabolic process	ACC, PRAD
GO:0008210	Estrogen metabolic process	BRCA, UCEC, OV
GO:0071870	Cellular response to catecholamine stimulus	PCPG, ACC
GO:0061621	Canonical glycolysis	LGG, HNSC, LUSC, ESCA, CESC, SKCM, LUAD
GO:0090141	Positive regulation of mitochondrial fission	LGG, SARC, SKCM, LUAD
GO:0000077	DNA damage checkpoint signaling	OV, SARC, BRCA, STAD, UCEC
GO:0043406	Positive regulation of MAP kinase activity	SKCM, LUAD, COAD, PAAD, LUSC, HNSC
GO:0042102	Positive regulation of T-cell proliferation	DLBC, THYM, HNSC, LUSC, SKCM
GO:0050671	Positive regulation of lymphocyte proliferation	DLBC, THYM, HNSC, BRCA
GO:0030198	Extracellular matrix organization	SARC, PAAD, BRCA, COAD, STAD
GO:0030199	Collagen fibril organization	SARC, PAAD, BRCA, COAD, STAD
GO:0010718	Positive regulation of epithelial-to-mesenchymal transition (EMT)	HNSC, SARC, BRCA, ESCA, CESC, SKCM
GO:0031643	Positive regulation of myelination	LGG
GO:0070572	Positive regulation of neuron projection regeneration	LGG
GO:0046951	Ketone-body biosynthetic process	LIHC, CHOL, KIRC, KIRP
GO:0030195	Negative regulation of blood coagulation	LIHC, CHOL

Table S2: Selection of 20 GO-terms of processes that are expected to vary in activity across cancer types. These terms are used to evaluate the interpretability of individual GONNECT node activations.

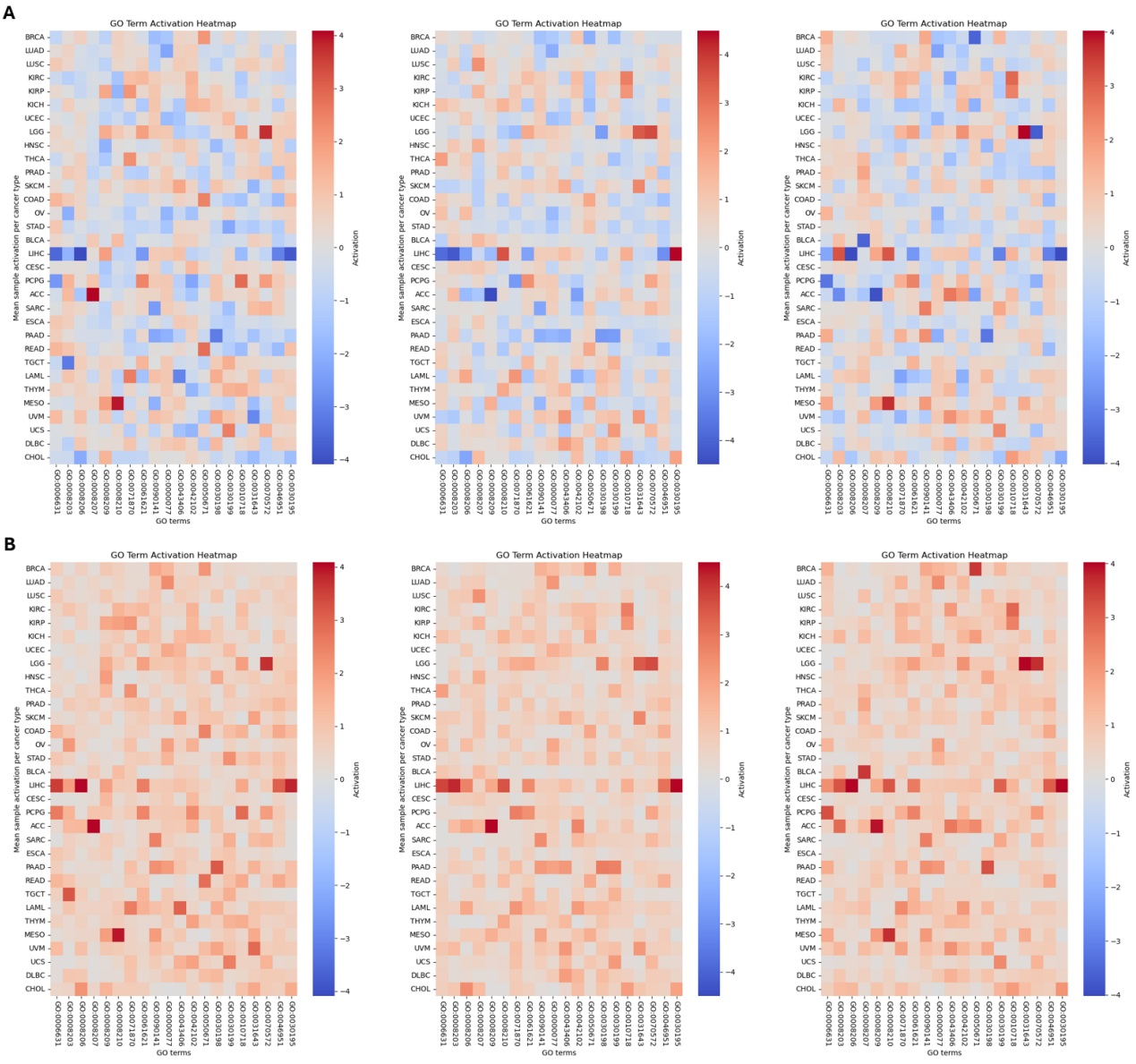


Figure S4: Mean activations per cancer type of the terms in Table S2. The three heatmaps per panel show three instances of a GONNECT encoder model. **A**) Raw activations per term, per cancer type. **B**) Absolute values of the mean activations in panel A. Panel A shows how the sign of the mean activations appears random across model instances. Panel B shows how the magnitude of some term-cancer type pairs are consistently high across different encoder instances.

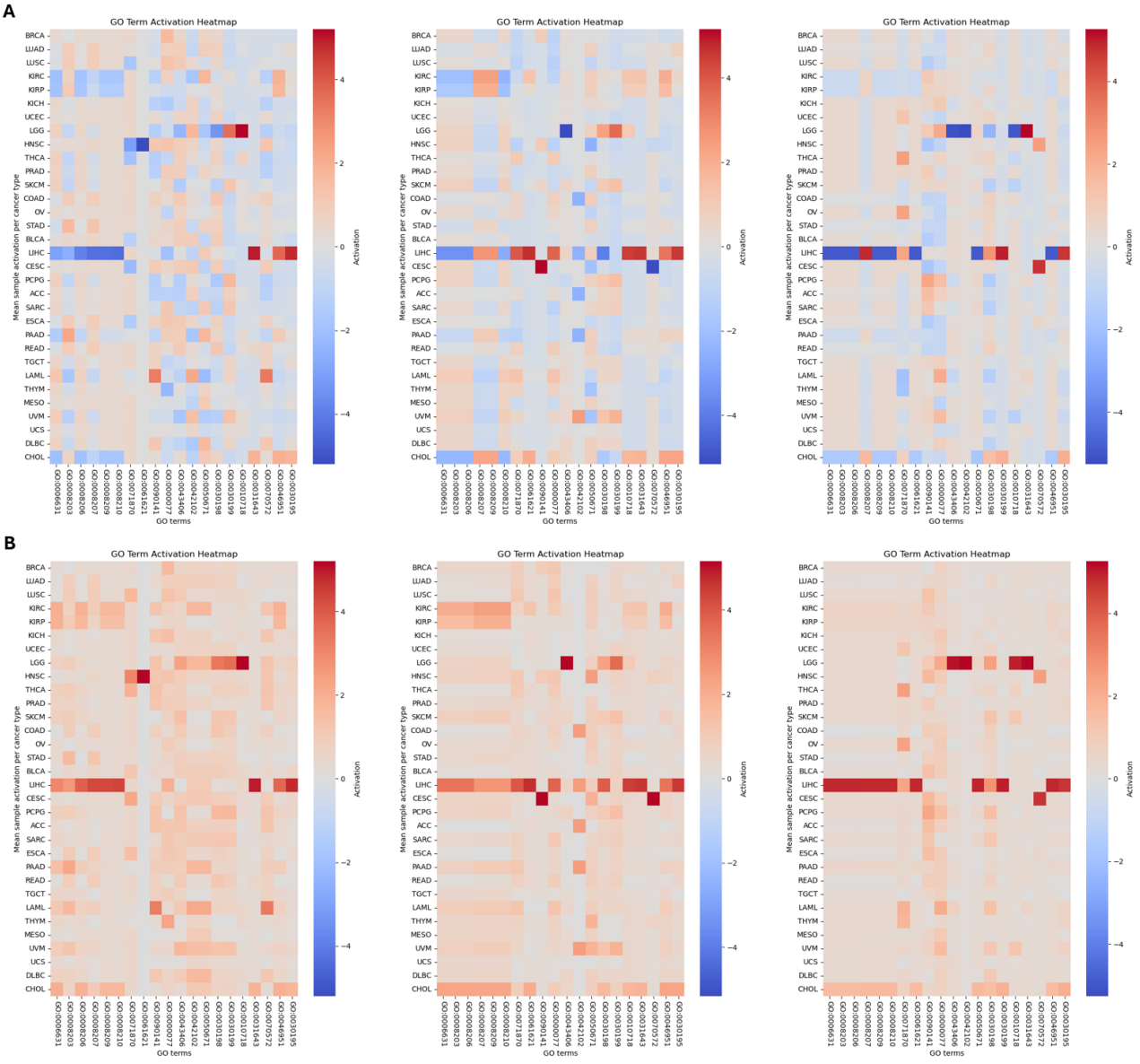


Figure S5: Mean activations per cancer type of the terms in Table S2. The three heatmaps per panel show three instances of a GONNECT decoder model. **A**) Raw activations per term, per cancer type. **B**) Absolute values of the mean activations in panel A. Panel A shows how the sign of the mean activations appears random across model instances. Panel B shows how the magnitude of some term-cancer type pairs are consistently high across different decoder instances.

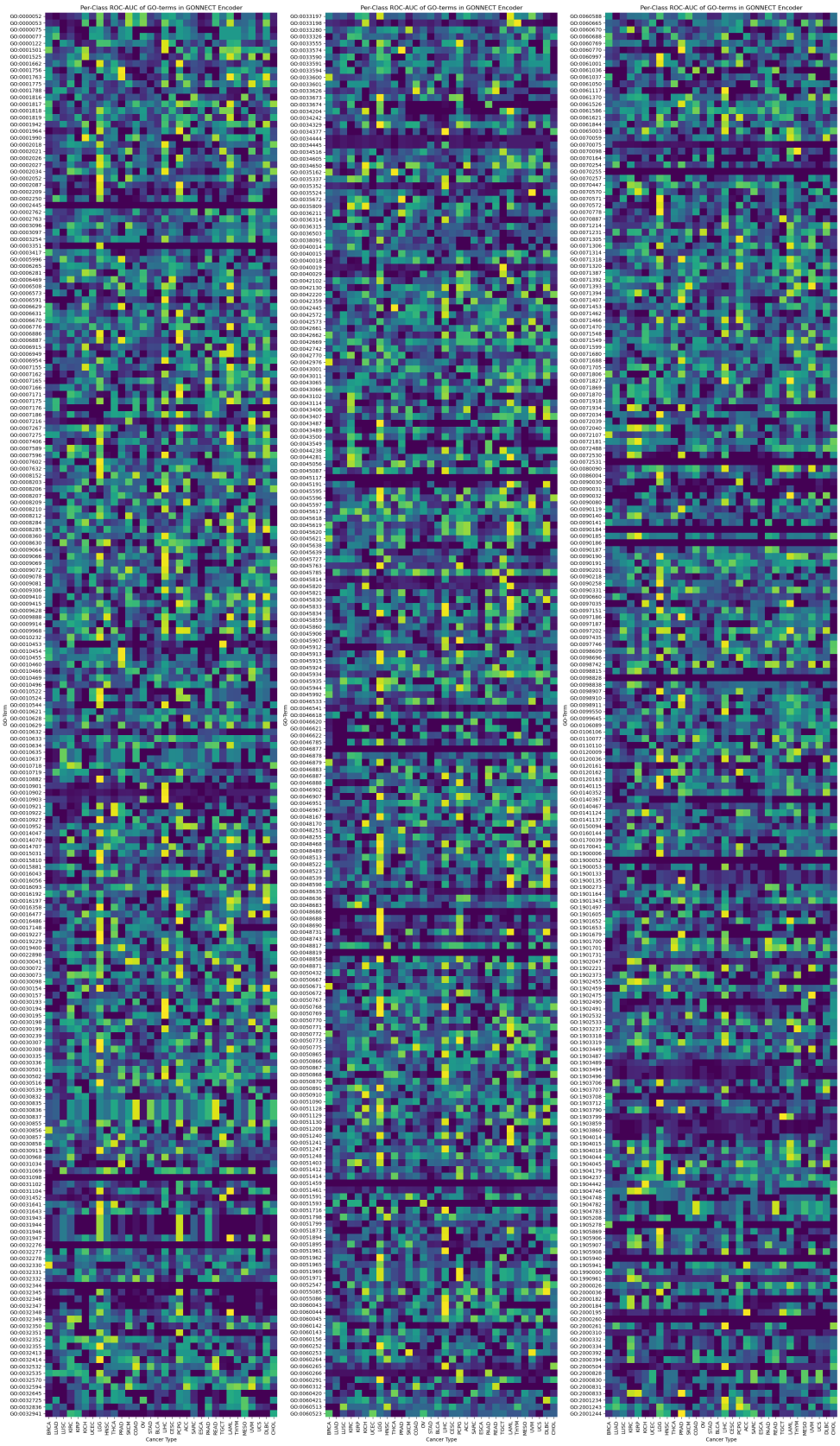


Figure S6: ROC-AUC scores per cancer type of all 623 GO-term associated nodes in the GONNECT encoder. The associated terms are denoted by GO ID and cancer types by their TCGA abbreviation.

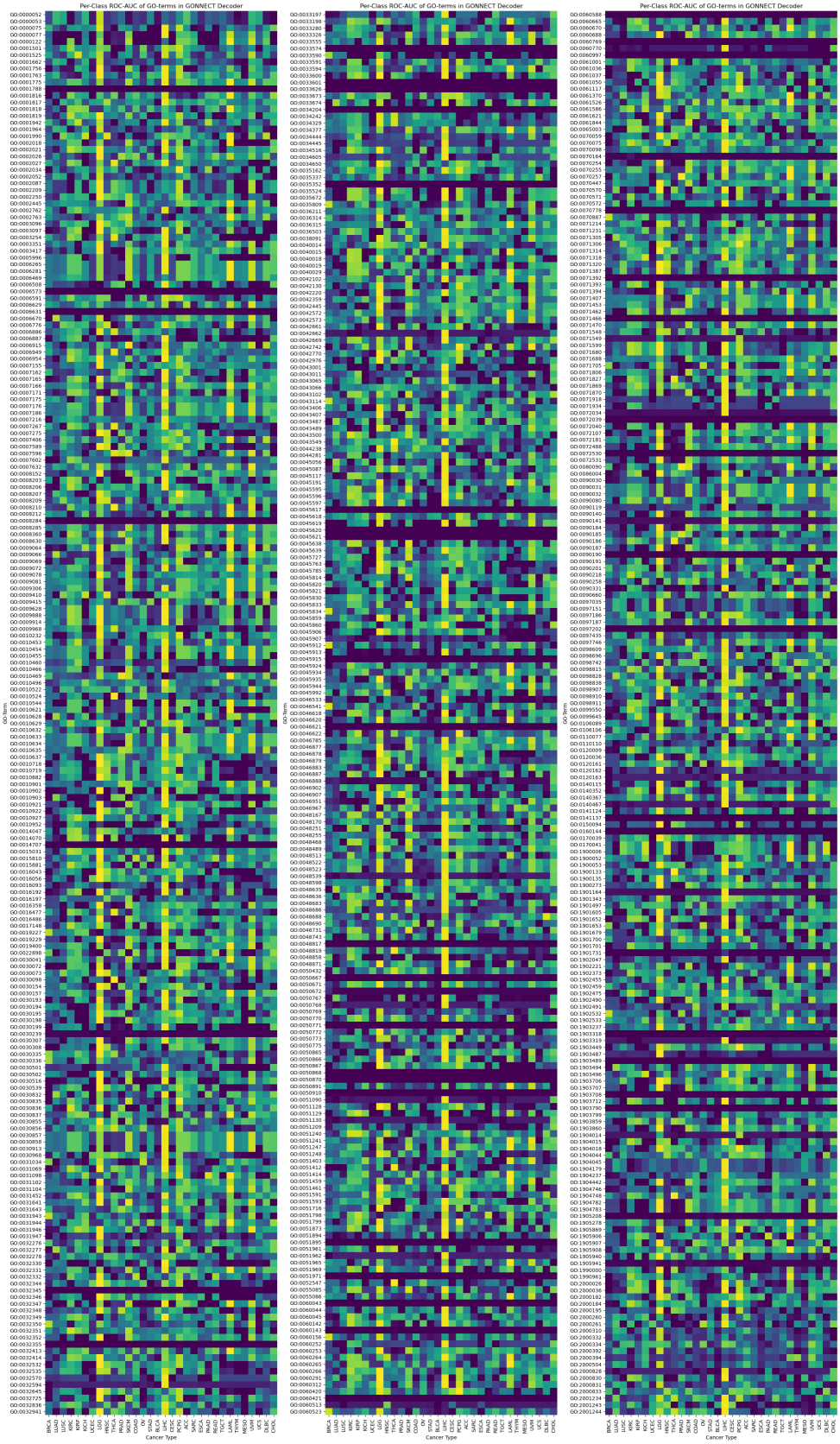


Figure S7: ROC-AUC scores per cancer type of all 623 GO-term associated nodes in the GONNECT decoder. The associated terms are denoted by GO ID and cancer types by their TCGA abbreviation.

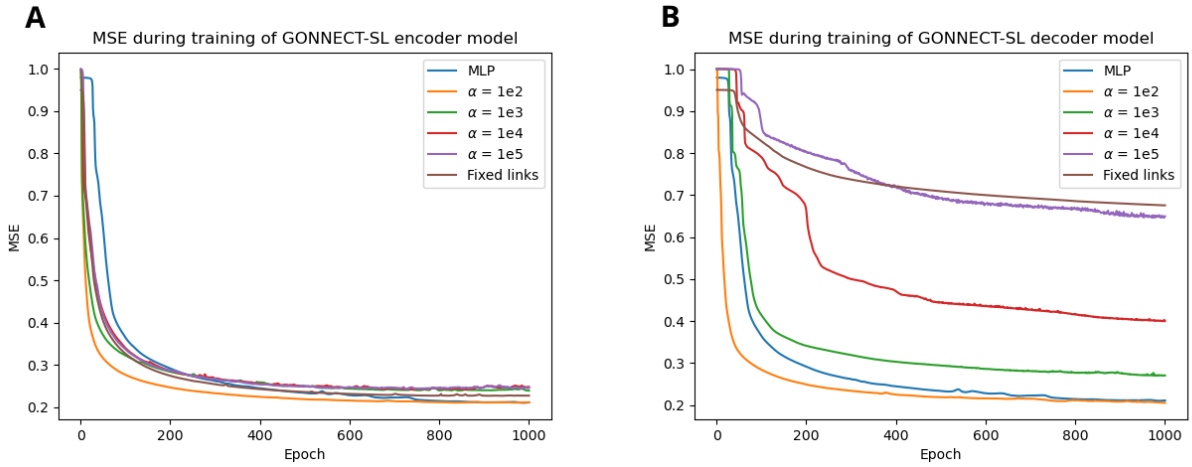


Figure S8: Loss curves of GONNECT-SL training with different values for hyperparameter α. The models are being trained with a specialized loss function, however, the figure shows regular mean square error (MSE) of input reconstruction. **A**) Loss curve of a GONNECT-SL encoder module. **B**) Loss curve of a GONNECT-SL decoder module. The larger α, the larger the weight regularization on soft links. A larger α results in less active soft links, which is favorable for interpretability, but harms reconstruction performance. The effect on reconstruction performance is more evident in the GONNECT-SL decoder compared to the encoder.

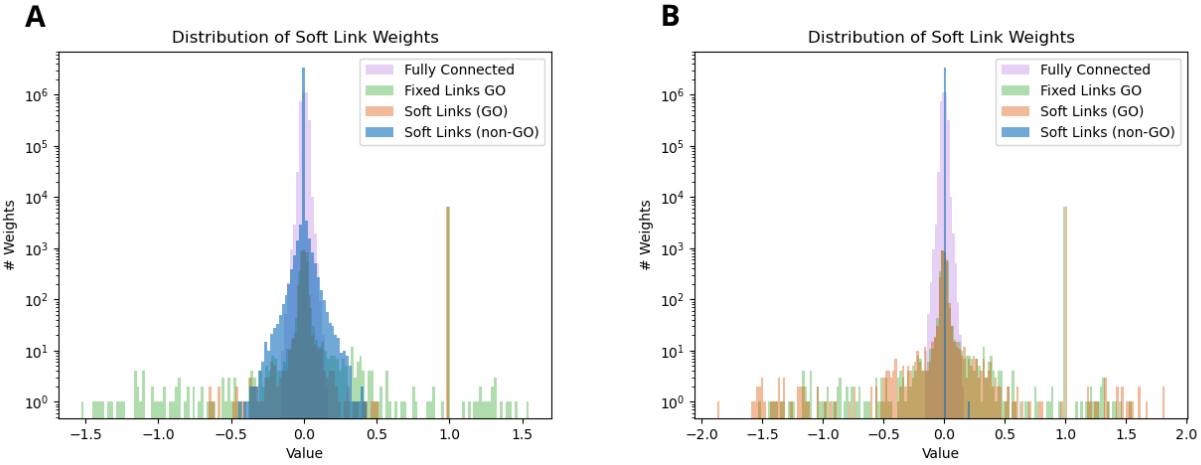


Figure S9: Weight distributions for different values of soft link hyperparameter α. For comparison, the weight distribution of a fully connected MLP (purple), and the distribution of regular GONNECT (green) is shown. The weights of GONNECT-SL are split up into the weights of links that were already present in GO (orange) and those that are not, meaning they can become active soft links (blue). **A**) Weight distribution for $\alpha = 1 * \cdot 10^2$. The relatively low value means that a lot of soft links are active, and most active soft links have relatively small magnitudes. Original GO links become less valuable, and therefore lose their high magnitude as well. **B**) Weight distribution for $\alpha = 1 \cdot 10^4$. The high value means that just one soft link becomes active, and the model essentially becomes equivalent to a fixed link GONNECT model.

Model	Time per epoch	Epochs required	Memory required	Total training time
MLP	1.1s	468	0.9 GB	0h 9m
GONNECT-SL enc	6.4s	501	1.1 GB	0h 53m
GONNECT-SL dec	6.1s	417	1.1 GB	0h 42m
GONNECT-SL both	10.0s	1166	1.1 GB	3h 14m
GONNECT enc	6.0s	462	1.1 GB	0h 46m
GONNECT dec	7.8s	2251	1.1 GB	4h 52m
GONNECT both	9.7s	8303	1.1 GB	22h 25m
GONNECT-R enc	4.6s	776	1.1 GB	0h 59m
GONNECT-R dec	6.4s	2163	1.1 GB	3h 51m
GONNECT-R both	8.1s	8700	1.1 GB	19h 39m

Table S3: Training statistics of the different model variants. Values in the table denote the means over five independent training runs on different data splits.

Supplementary References

[1] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000. DOI: 10.1038/75556.

[2] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (tcga): An immeasurable source of knowledge," *Contemporary Oncology/Współczesna Onkologia*, pp. 68–77, 2015, ISSN: 1428-2526. DOI: 10.5114/wo.2014.47136.

[3] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(87)80084-9.

[4] L. McInnes, J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, 2020. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426 [stat.ML].