Studying a disease through the lens of its related complex data

Marta Kwiatkowska
Weronika Pędzimąż
David Riley

Medical terminologies and semantic datasets

1.1. Disease general information.

Non-Small Cell Lung Cancer (NSCLC) is the most common form of lung cancer and the leading cause of lung cancer deaths, defined by excluding small-cell lung carcinomas. The primary cause is tobacco smoking, accounting for 90% of cases, though exposure to asbestos, radon, pollution and radiation also contribute. Symptoms often include persistent cough, chest pain, shortness of breath, coughing up blood, and hoarseness, with potential spread to bones or the brain in advanced stages.

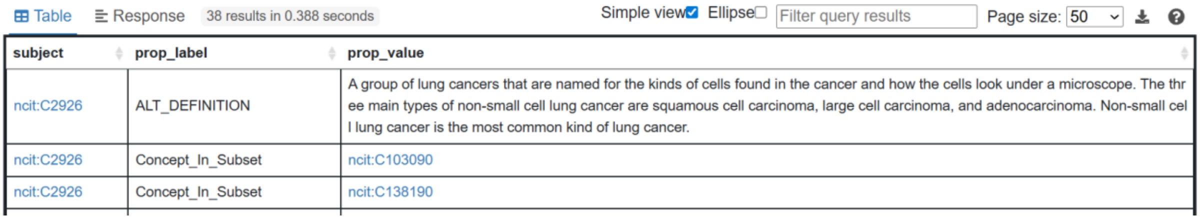
Diagnosis typically begins with a chest X-ray and CT scan, followed by a biopsy for confirmation, PET scans and brain MRIs may be used to assess spread. Treatment varies by stage: early-stage disease is often treated with surgery, potentially followed by chemotherapy or radiation, while advanced disease may require chemotherapy, targeted drugs, or immunotherapy like pembrolizumab.

Codifications for Non-Small Cell Lung Carcinoma (NSCLC):

- UMLS CUI: C0007131,
- MeSH: D002289,
- ICD-10: NSCLC cases are coded under “Malignant neoplasm of bronchus and lung” (C34.0–C34.9), there is no unique subtype code is assigned for NSCLC,
- SNOMED CT: 254637007 (Non-small cell lung cancer),
- NCIt (NCI Thesaurus): C2926.

1.2. SPARQL Query and Summary

The execution of a SPARQL query on the NCIt database for concept ncit:C2926 successfully retrieved 38 annotation properties, providing semantic context for the disease. The results highlighted a detailed alternative definition that identifies the condition as the most common form of lung cancer and explicitly categorizes its three main subtypes: squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. Additionally, the data revealed the concept's hierarchical organization through its inclusion in specific subsets, such as ncit:C103090.



subject	prop_label	prop_value
ncit:C2926	ALT_DEFINITION	A group of lung cancers that are named for the kinds of cells found in the cancer and how the cells look under a microscope. The three main types of non-small cell lung cancer are squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. Non-small cell lung cancer is the most common kind of lung cancer.
ncit:C2926	Concept_In_Subset	ncit:C103090
ncit:C2926	Concept_In_Subset	ncit:C138190

Bioinformatics

2.1. Disease genes

ALK (Anaplastic Lymphoma Kinase):
The ALK gene provides instructions for making a protein that supports cell growth and development. In some lung cancers, ALK becomes abnormally rearranged, creating a faulty protein that drives cancer cell growth. Targeted medicines can block this abnormal ALK activity. NCBI Gene ID: 238, NCBI Reference Sequence: NG_009445.1

FASTA file (header and first two lines):

```
>NG_009445.1:4956-733793 Homo sapiens ALK receptor tyrosine kinase (ALK), RefSeqGene
(LRG_488) on chromosome 2
AGCTGCAAGTGGCGGGCGCCCAGGCAGATGCGATCCAGCGGCTCTGGGGGCGGCAGCGGTGGTAGCAGCT
GGTACCTCCCGCCGCCTCTGTTCGGAGGGTCGCGGGGCACCGAGGTGCTTTCCGGCCGCCCTCTGGTCGG
```

EGFR (Epidermal Growth Factor Receptor) produces a cell surface protein for growth signaling; mutations lead to continuous division and tumor progression. KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog) regulates cell communication; mutations keep growth signals active, causing uncontrolled multiplication.

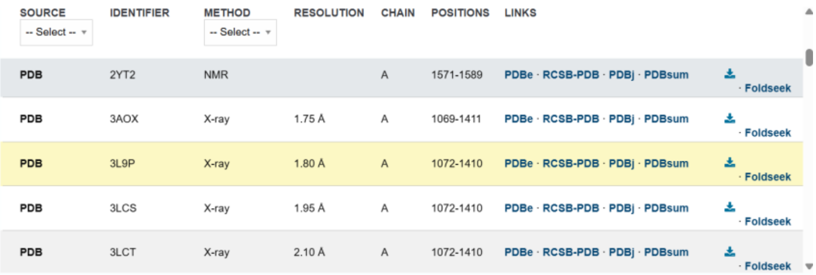
2.2. Proteins.

Following gene identification, the ALK protein (UniProt Q9UM73) was analyzed. ALK is a receptor tyrosine kinase involved in nervous system development. Pathological EML4-ALK fusion leads to ligand-independent activation and carcinoma.

Protein sequence: (FASTA file, header with first two lines)

```
>sp|Q9UM73|ALK_HUMAN ALK tyrosine kinase receptor OS=Homo sapiens OX=9606 GN=ALK PE=1
SV=3
MGAIGLLWLLPLLLSTAAVGSGMGTGQRAGSPAAGPPLQPREPLSYSRLQRKSLAVDFVV
PSLFRVYARDLLLPPSSSELKAGRPEARGSLALDCAPLLRLLGPAPGVSWTAGSPAPAEA
```

The AlphaFold structure AF-Q9UM73-F1-model_v4.pdb was utilized. Verification via UniProt (Q9UM73) confirms experimental structures exist (e.g., 3AOX, 2YT2) but are limited to domains like the intracellular kinase region rather than the full protein.



SOURCE -- Select --	IDENTIFIER	METHOD -- Select --	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	2YT2	NMR		A	1571-1589	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	3AOX	X-ray	1.75 Å	A	1069-1411	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	3L9P	X-ray	1.80 Å	A	1072-1410	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	3LCS	X-ray	1.95 Å	A	1072-1410	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	3LCT	X-ray	2.10 Å	A	1072-1410	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek

Figure 1 - List of experimental structures available in UniProt, showing X-ray and NMR methods covering only partial positions

Experimental structures offer high-resolution views (1.75–2.50 Å) but fail to resolve the full-length protein, often omitting disordered extracellular regions and linkers. In contrast, the AlphaFold model provides a complete 3D representation (residues 1–1620), successfully predicting the parts missing from classical X-ray structures.

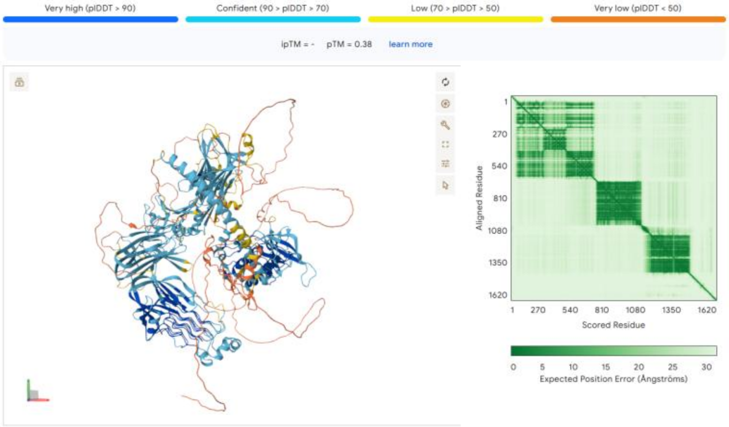


Figure 2 - AlphaFold model confidence (pLDDT) and alignment error.

A comparison of the two approaches reveals that the AlphaFold model aligns closely with the X-ray structure in the kinase domain with high confidence, while previously missing N-terminal and linker regions are included but show low confidence, suggesting a disordered state. To reflect the dynamic nature of the ALK receptor, the AlphaFold query was updated to include Adenosine Triphosphate, transitioning the model from a static apo-structure to an active holo-structure. This generated a realistic 3D visualization of the biologically active site where therapeutic drugs compete with ATP to inhibit the cancer signaling pathway.

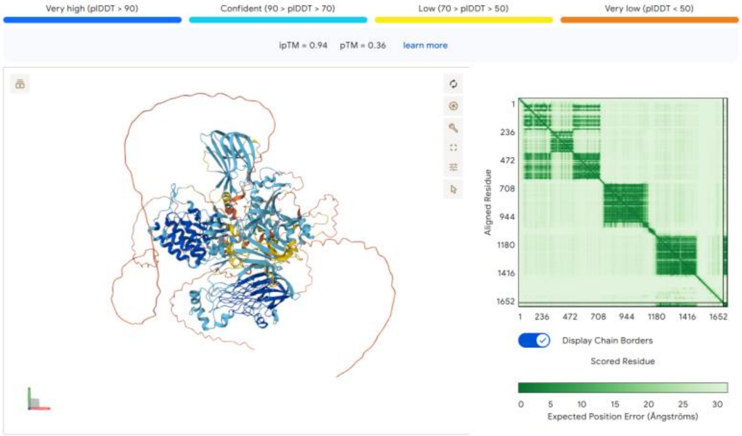
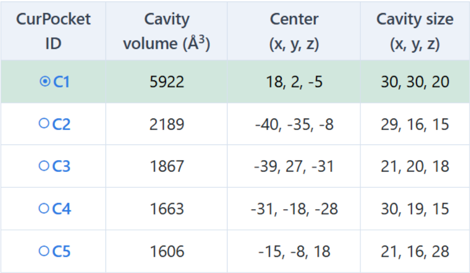


Figure 3 - Model ALK with ATP, obtained using AlphaFold Server

To identify precise therapeutic binding sites, a blind cavity detection analysis was performed using the CB-Dock2 server on the AlphaFold PDB structure with ATP and ions removed. The server's geometry-based algorithm scanned the protein surface, successfully identifying the optimal binding cavity within the intracellular kinase domain. The resulting center coordinates and box sizes provide a specific 3D map to guide future molecular docking experiments and inhibitor design.



CurPocket ID	Cavity volume (Å^3)	Center (x, y, z)	Cavity size (x, y, z)
◉C1	5922	18, 2, -5	30, 30, 20
○C2	2189	-40, -35, -8	29, 16, 15
○C3	1867	-39, 27, -31	21, 20, 18
○C4	1663	-31, -18, -28	30, 19, 15
○C5	1606	-15, -8, 18	21, 16, 28

Figure 4 - Table of results from CB-Dock2 showing the top predicted binding cavities.

The analysis successfully identified the best cavity (C1) located within the intracellular kinase domain. According to the results, Cavity 1 is located at center coordinates (18, 2, -5) with a box size of 30, 30, 20 Å.

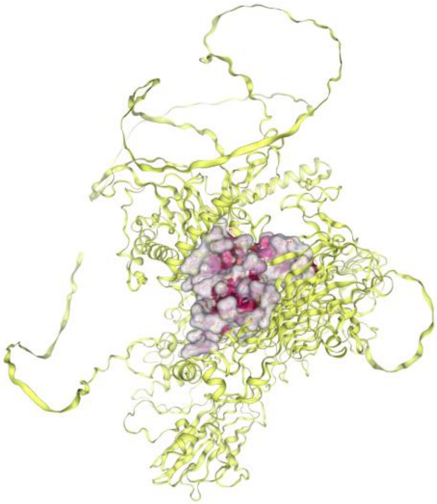


Figure 5 - Visualization of the protein with the predicted binding pocket (C1) highlighted.

2.3. Transcriptome.

To investigate the transcriptional consequences of *IGF2BP3* silencing in lung cancer cells, we analyzed RNA-Seq data from the GSE298476 dataset, comprising four samples: two biological replicates with *IGF2BP3* knockdown and two control replicates. Raw count data were processed using a standard bioinformatics pipeline in R, beginning with quality control and filtration to remove low-expression noise, which reduced the dataset from 20,263 to 13,379 analyzable genes. Normalization was performed using the TMM method (Trimmed Mean of M-values), followed by precision weighting via the voom transformation to enable linear modeling with the *limma* package. Differential expression analysis revealed a moderate transcriptional response; strict filtering (FDR < 0.05) yielded limited targets, necessitating a relaxed criteria (P-value < 0.01) which successfully identified 110 differentially expressed genes (DEGs). These DEGs were subjected to Gene Ontology (GO) enrichment analysis to uncover potential biological implications of the knockdown. The expression patterns of the top candidates distinguish the knockdown condition from the control, confirming the experimental validity.

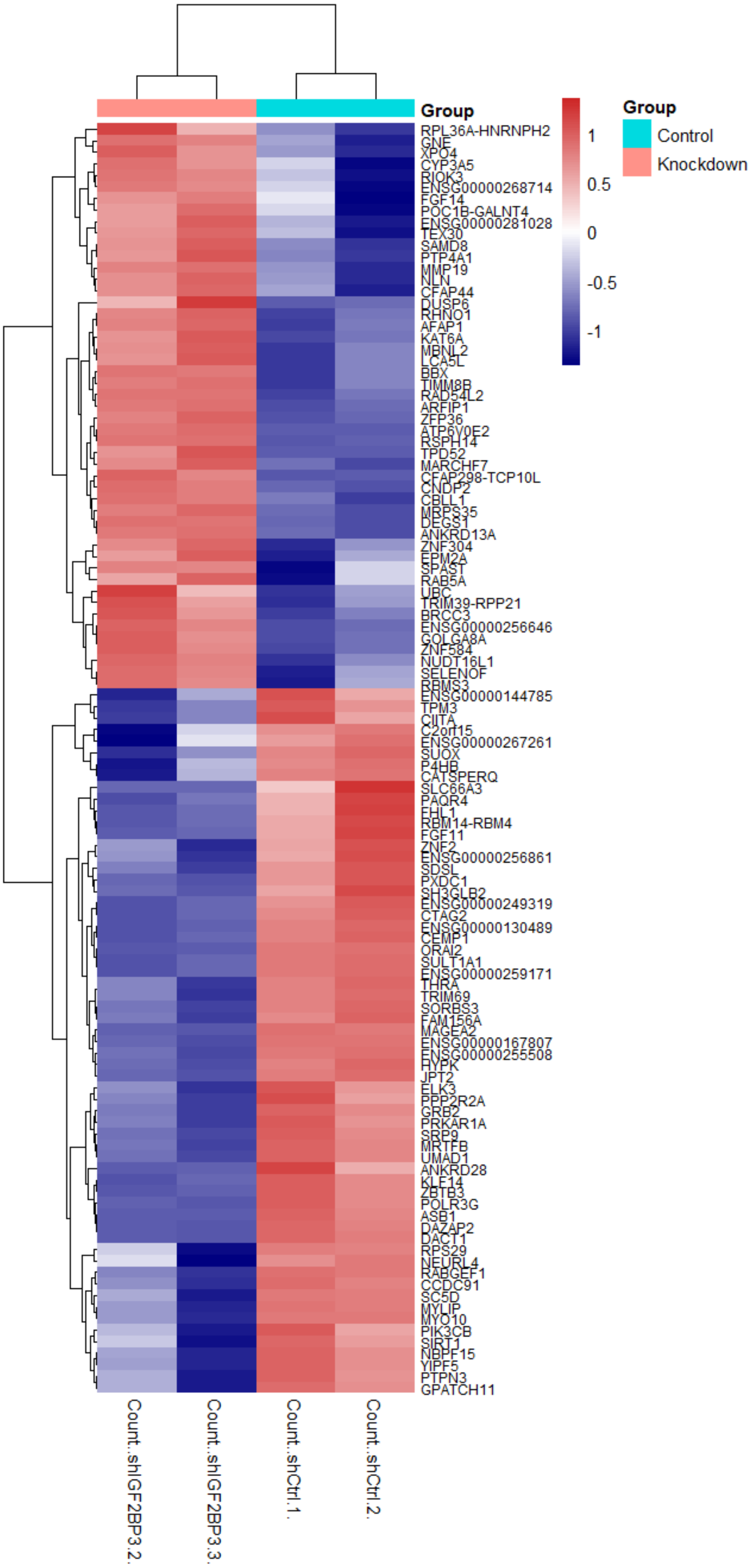


Figure 6 - Heatmap of top differentially expressed genes between IGF2BP3 knockdown and control samples.

At Figure 6, rows represent individual genes, while columns represent biological replicates . Color intensity indicates relative expression levels (Z-score), with red denoting upregulation and blue denoting downregulation relative to the mean. The clear clustering of samples by condition validates the reproducibility of the biological replicates.

The transcriptomic analysis of *IGF2BP3* knockdown identified a discrete signature of 110 differentially expressed genes, indicating a specific rather than global regulatory role. Distinct clustering of samples confirms the experimental validity and reproducibility of the shRNA treatment. These findings suggest that *IGF2BP3* modulates a select subset of transcripts in NSCLC cells, providing a targeted list of potential downstream effectors for further investigation.

Network medicine

3.1. Disease module.

The primary objective of this analysis was to validate the "Disease Module Hypothesis" for Non-Small Cell Lung Carcinoma (NSCLC), which suggests that disease-associated genes cluster in specific topological neighborhoods within the human interactome. A protein-protein interaction (PPI) graph was constructed using NetworkX, with self-loops removed to ensure accurate topology. Disease genes were identified from the *disease_gene.tsv* database using the standardized name 'non-small cell lung carcinoma' (UMLS C0007131), filtering out non-specific entries. This yielded 156 associated genes, 146 of which were successfully mapped to the constructed interactome.

Analysis of the subgraph induced by these 146 genes revealed a Largest Connected Component (LCC) of 109 genes, indicating that approximately 75% of the mapped genes interact directly to form a distinct module. To validate statistical significance, a Monte Carlo simulation with 1,000 permutations of 146 randomly selected genes was performed. The random sets produced a mean LCC size of approximately 9.8 genes, whereas the observed NSCLC LCC of 109 resulted in a Z-score of 13.91 (p-value ≈ 0.0), confirming the non-random clustering of these genes.

The findings validate the topological localization of NSCLC within the interactome. A significance histogram was generated to visualize the deviation from random expectation, and the identified 109-gene module was exported as a .gexf file (*NSCLC_disease_module.gexf*) for further visualization and analysis of hub proteins in software such as Gephi.

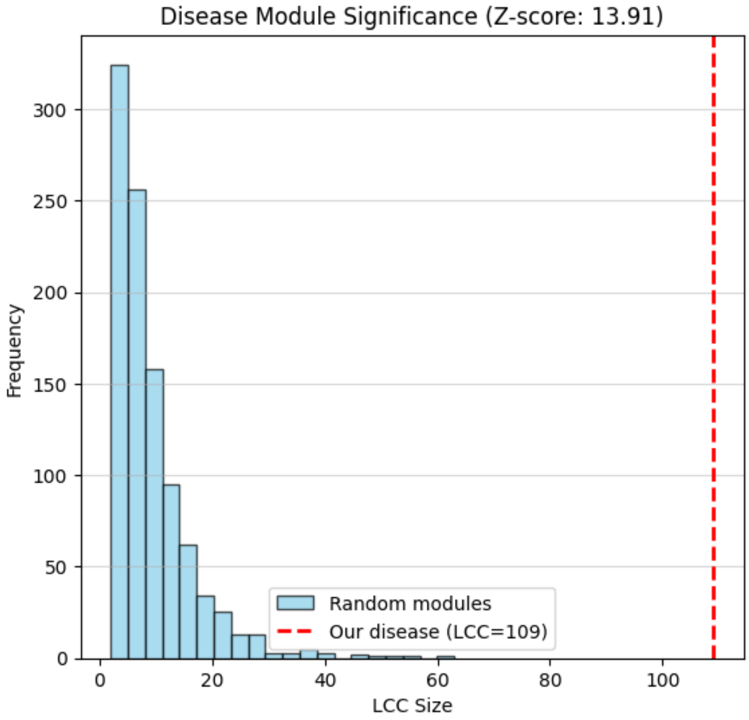


Figure 7 - Significance histogram, the statistical distance between our result and the random expectation.

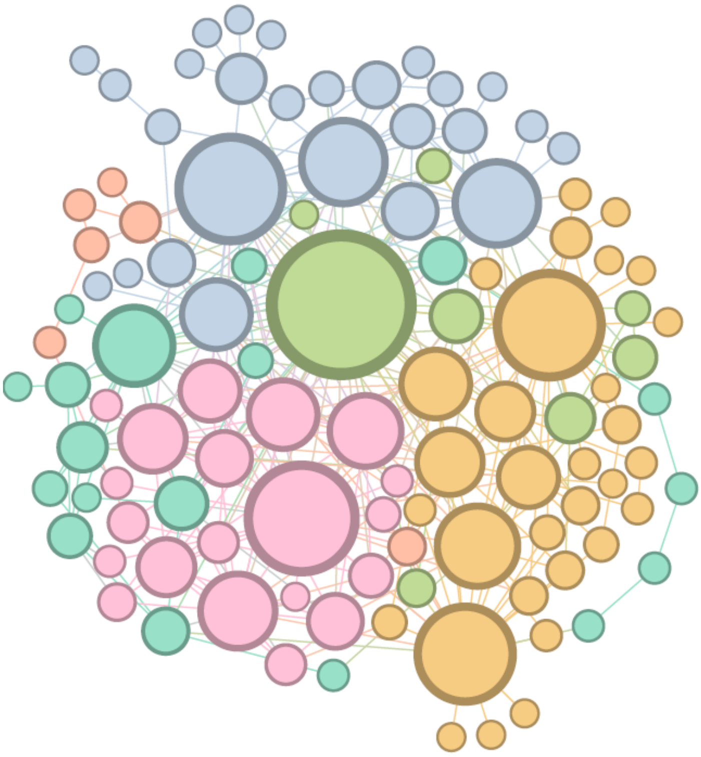


Figure 8 - Graph formed by the disease module in Gephi

3.2. Disease separation.

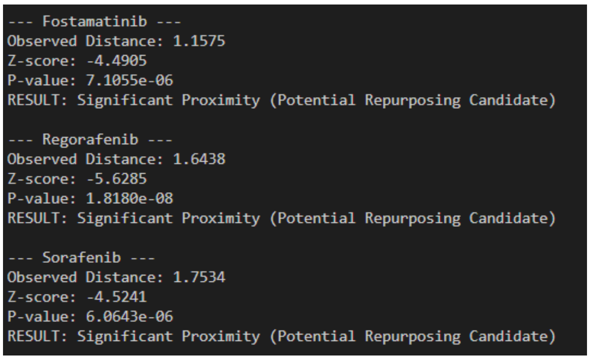
The objective of this analysis was to place the Non-Small Cell Lung Carcinoma (NSCLC) module within the human interactome by quantifying its topological relationship with other disease modules. According to the network medicine framework, biologically related diseases are expected to show overlapping or proximal modules, whereas unrelated phenotypes should be topologically separated. To evaluate this, network separation was calculated between NSCLC and two reference diseases: Small Cell Lung Carcinoma (SCLC), chosen due to shared tissue of origin and risk factors, and Schizophrenia, selected as a biologically distant disorder.

The separation between NSCLC and SCLC (146 genes each) was ~0.0000, indicating an almost overlapping but technically distinct network relationship. This near-zero value suggests that the two lung cancer modules are immediate network neighbors, consistent with their classification as distinct histological subtypes that share core oncogenic pathways. In contrast, the separation between NSCLC and Schizophrenia (851 genes) was 1.4106, demonstrating strong topological separation and largely independent molecular interaction neighborhoods. Together, these results confirm that the network separation metric effectively distinguishes between biologically related and unrelated disease modules.

3.3. Disease - drug proximity.

Apart from the drugs that are specifically indicated for the disease you are studying, can you provide some new repurposing opportunities? Look at some of the drugs whose targets are specifically in the module of your disease and use the proximity metric to describe how near the disease and the drug are.

This analysis aimed to identify potential therapeutic candidates by evaluating the topological proximity between drug targets and the Non-Small Cell Lung Carcinoma (NSCLC) disease module. We screened the drug-target database and identified 147 drugs with multiple targets residing directly within the NSCLC gene module. We selected the top candidates based on target overlap, Fostamatinib, Regorafenib, and Sorafenib, and computed their network proximity relative to random expectation.



```
--- Fostamatinib ---
Observed Distance: 1.1575
Z-score: -4.4905
P-value: 7.1055e-06
RESULT: Significant Proximity (Potential Repurposing Candidate)

--- Regorafenib ---
Observed Distance: 1.6438
Z-score: -5.6285
P-value: 1.8180e-08
RESULT: Significant Proximity (Potential Repurposing Candidate)

--- Sorafenib ---
Observed Distance: 1.7534
Z-score: -4.5241
P-value: 6.0643e-06
RESULT: Significant Proximity (Potential Repurposing Candidate)
```

All three drugs exhibited highly significant negative z-scores, confirming that their targets are topologically closer to the NSCLC module than expected by chance (z << 1.5). Regorafenib demonstrated the strongest network proximity, targeting key nodes such as *KDR*, *RET*, and *BRAF*. Sorafenib and Fostamatinib also showed strong significance. These results validate the network medicine hypothesis for these compounds, suggesting that their efficacy stems from modulating the specific local neighborhood of the interactome associated with lung carcinoma.

Medical images

4.1. Task selection

We used the IQ-OTH/NCCD lung cancer dataset from Kaggle (Al-Yasriy, 2020). It contains approximately 1100 X-ray images, labelled as normal, malignant or benign cases. Because the image labels are categorical, we determined that the appropriate task for this dataset is classification. As the task description was rather open-ended, we decided to train 2 separate models and compare their performance, as we thought that it would give us more practical experience and insight compared to just training a single one. We settled on fine-tuning the classifier of vgg16 and to design our own simpler model. It is also worth noting that this was an iterative process and will thus be described as such in future sections. This section will be referring to the submitted jupyter notebook as it was considered easier to follow for readers. The notebook sections map directly to the report sections.

4.2. Data preparation

As previously mentioned, the dataset contained 3 categories (labels). The frequency count can be seen in the table below.

Label	Benign	Normal	Malignant
Count	416	120	561

Easily noted is that it is imbalanced and thus we provided our loss function with class weights computed as (CF denoting the set of class frequencies):

$$\left\{\frac{avg}{f} : f \in CF\right\}, avg = \frac{1}{3}\sum_{i=1}^{3} \quad f_i$$

The images were of varying sizes too, the vast majority were uniformly sized. Since we resized the images that didn't conform to the majority, it's possible that it may have degraded their quality somewhat. However we still considered manual cropping as a potential future micro-optimization that we may do in a real-world scenario. The frequency count for the image dimensions can be seen below.

Dimensions	512 x 512	511 x 404	801 x 512	506 x 331
Count	1036	1	31	1

Because our 2 models had different input dimensions, the preprocessing of the dataset was slightly different between the two. For vgg16, the model input expects a 3 channel 224x224 image of the same colour distribution as its training data (imagenet). This required a resize, and normalization transformation as can be seen in 4.2 of the provided notebook. Our model was designed to accept grayscale images directly, and thus requires reducing the number of input colour channels from the native 3 channel png files to 1 channel tensors. It would also be unnecessarily computationally expensive to work with the native 512x512 resolution and we thus decided to take 256x256 as the input dimension. The models are further discussed in section 4.3.

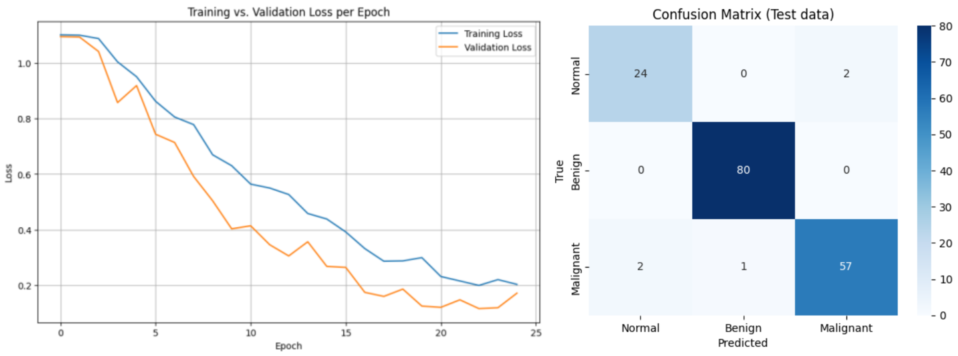
The data was then split into 3 random partitions with percentages 70/15/15 for training, validation and testing respectively. Even though the task description never explicitly mentioned a validation partition, we included it anyway as it seems to be industry standard to do so and provides valuable insights into the training process. Thanks to it, we could easily discover overfitting and errors in our designs. It also serves as a great metric to conditionally terminate training based on model performance at any given epoch. The random split was seeded with the same seed for both models to allow for a fair comparison.

4.3 Architecture selection and model training

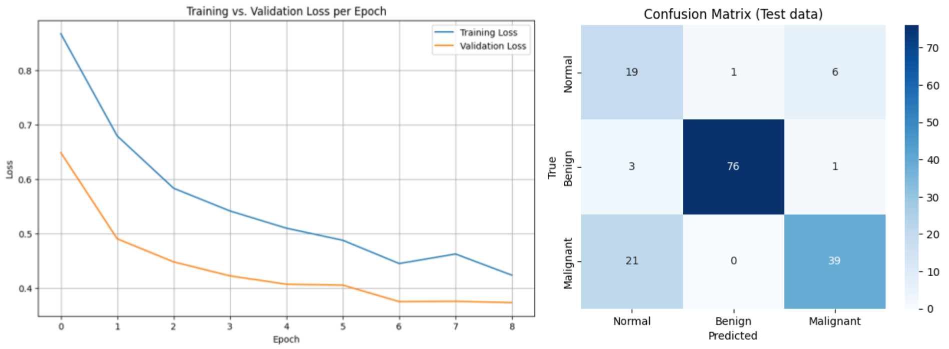
We chose to both fine-tune the vgg16 model shown in class, using pretrained *imagenet* weights and to design our own simpler classification model using random weights. The detailed description of the architectures can be seen in 4.3 of the notebook. The only notable change to vgg16, was changing the last linear layer to accommodate 3 classes. Our simpler model, hereby named SimpleLCC (Lung Cancer Classifier), went through several iterations during experimentation, the added changes are marked with comments in the notebook and were added to counteract excessive overfitting. These include an adaptive average pooling layer in the feature section and a dropout layer in the classifier.

For a fair comparison, the two models used the same built-in pytorch Adam (Adaptive Moment Estimation) optimizer. However, we found that using different learning rates: $8 \cdot 10^{-3}$ for SimpleLCC and $3 \cdot 10^{-4}$ for vgg16 made them converge at somewhat similar rates and ended up being the only differentiating factor during training. Both also used the same pytorch Cross Entropy Loss function providing the label weights discussed in 4.2.

The training loop used was very similar to that of the teacher's example. Some additional early termination conditions were added during early model development, as loss would explode at various times due to bugs we ourselves introduced. We iterated with 10 epochs, and ended up using 50 for the final training. However the training loops exited early in all cases, as can be seen on the x-axes of the graphs. Luckily for us, one of us had a CUDA device available, which rapidly sped up the training loops, allowing for more iteration. The training metric graphs for the final two versions of the models can be seen below. We also experimented with different batch sizes which had an effect on the accuracy and especially malignant recall. It ranged between 88-98% across batch sizes 3-16 for SimpleLCC, indicating sensitivity to training stochasticity given the limited dataset size. For a less noisy result and smoother training, we chose to keep a batch size of 16 consistent for both models during the final training. It is worth noting however that a high malignant recall may be the most preferable metric of all and so batch size choice matters a lot in that regard.



Training metrics and confusion matrix for SimpleLCC



Training metrics and confusion matrix for VGGLCC

4.4 Model evaluation

We used our test partition not previously seen by either model to evaluate the performance of both after training. We used the *classification_report* and *confusion_matrix* functions from the sklearn package to obtain our metrics. They can be seen below.

	support	precision		recall		f1-score	
		SimpleLCC	**VGG16**	**SimpleLCC**	**VGG16**	**SimpleLCC**	**VGG16**
Normal	26	0.92	0.44	0.92	0.73	0.92	0.55
Benign	80	0.99	0.99	1.00	0.95	0.99	0.97
Malignant	60	0.97	0.85	0.95	0.65	0.96	0.74
Accuracy		**SimpleLCC**	**VGG16**				
		0.97	0.81				

In general, we noticed a significant difference between the models during both training and testing. SimpleLCC is a much smaller model, almost 94 times smaller, and so it completed training faster. Despite using a pretrained VGG16 model, performance was inferior to a smaller custom CNN. We think that this is likely due to the domain mismatch between the *ImageNet* dataset images consisting of coloured natural images and the medical dataset with grayscale X-ray imagery. Since the whole convolutional block was frozen, it is not particularly surprising that the model was likely to perform a lot worse. This also explains why it has very imbalanced metrics, as the CNN layer was never optimized with class weights. A future improvement may be to unfreeze some of the later convolutional layers in the VGG16 model. However, the main bottleneck for our results is very likely the dataset size. It is unfortunately very small and quite an imbalanced one at that. We saw severe gradient noise for small batches and our early iterations on custom models with more parameters were quick to overfit the small dataset. We don’t feel very confident in trusting a model with potential higher recall for malignant cases with a support of 60 and can conclude the following from the assignment:

- The dataset is too small and unbalanced, making it unrealistic to expect a model trained on it to perform well in a real-world scenario
- For a smaller medical dataset, the use of a pretrained model is not necessarily better despite the added generalization of pretrained weights, if the domains of input aren’t themselves very similar.
- It is likely better to spend the effort augmenting the dataset and keep using a simpler model, than trying to optimize the model further.