Supplementary Material

Functional networks inference from machine learning models

Nicola Lazzarini, Paweł Widera, Stuart Williamson, Rakesh Heer, Natalio Krasnogor and Jaume Bacardit

1 Classification accuracy under feature selection

To choose the default percentage of attributes retained in the feature selection procedure, we performed a preliminary analysis using all 8 transcriptomic datasets from the main article. We evaluated how the BioHEL classification accuracy changes with the number of selected features (using linear SVM-RFE). The accuracy was measured using a standard 10 cross-fold validation. The full experiment (not reported here) used 100%, 90%, 80%, ..., 10% of the original dataset attributes.

We found that even when only 10% of the attributes are retained, the classification accuracy remains almost unchanged. Specifically, with 10% of the original attributes the accuracy increased for 2 datasets, slightly decreased for 3 datasets and remained unaltered for the other datasets. The exact results are reported in Table S1 below:

Table S1: BioHEL classification accuracy for each dataset, in 10-fold cross-validation experiments on the original and reduced set of attributes (before and after the feature selection). Linear SVM-RFE was used to select best 10% of the attributes.

| dataset       | all attributes | 10% attributes |
|---------------|---------------|----------------|
| Dlbcl         | 0.871         | 0.886          |
| CNS           | 0.473         | 0.451          |
| Leukemia      | 0.945         | 0.945          |
| Lung-Michigan | 0.980         | 0.980          |
| Lung-Harvard  | 0.978         | 0.964          |
| Prostate      | 0.892         | 0.892          |
| AML           | 0.637         | 0.592          |
| Colon-Breast  | 0.903         | 0.940          |

Given that we were able to maintain good classification accuracy despite large reduction in number of used attributes, we decided to use 10% attributes as a default setting for the FuNeL feature selection procedure. However, this FuNeL parameter is under the user control and the default setting can be changed.

2 Time complexity

The FuNeL protocol has four stages (see Figure 2 in the main article): (1) feature selection (optional), (2) rule-based network generation, (3) permutation test and (4) second rule-based network generation (optional).

The running time for the whole pipeline depends on the rule set generation time (execution time of BioHEL), as the optional feature selection stage can be seen as running in constant time. Two main factors that influence the rule set generation time are: (1) the number of attributes and (2) the number of samples.

We performed an execution time analysis of BioHEL using the largest (in terms of number of attributes) Colon-Breast dataset [Chowdary et al. 2006]. In the feature selection stage we retained: 20, 200, 2000, 10 000 and 20 000 attributes. From each of these 5 datasets we generated 100 random subsets of 50, 40, 30, 20 and 10 samples. Finally, we ran BioHEL 1000 times to obtain 1000 rule sets for each dataset.

Figure S1 shows the running times averaged across 100 000 runs (1000 runs for each of the 100 datasets).

The total execution time of FuNeL configurations $C_1$ and $C_2$ is calculated as:

$$T_1 = (\text{rule_sets} \times t(atts_1, \text{samples})) + (\text{permutation_runs} \times t(atts_1, \text{samples}))$$

(1)
Figure S1: Average execution times of a single BioHEL run for a given number of samples and attributes.

where \( \text{rule\_sets} \) is the number of inferred rule sets, \( \text{permutation\_runs} \) is the number of randomised datasets used in the permutation test and \( t(\text{atts}_1, \text{samples}) \) represents execution time of a single BioHEL run, that linearly depends on the size of a dataset measured in number of attributes and samples.

Configurations \( C_3 \) and \( C_4 \) require an additional run of BioHEL (step 4), and their total execution time is:

\[
T_2 = T_1 + (\text{rule\_sets} \times t(\text{atts}_2, \text{samples})) \tag{2}
\]

where \( \text{atts}_2 \) is the number of attributes after the permutation test (\( \text{atts}_1 \leq \text{atts}_2 \)).

It is important to notice that each run of BioHEL is independent, thus the generation of the rule sets can be trivially parallelised without any extra overhead. Given \( n \) computational cores, the total execution times could be reduced to:

\[
T_{\text{real1}} = \frac{T_1}{n} \quad T_{\text{real2}} = \frac{T_2}{n} \tag{3}
\]
3 Comparison of networks topological properties

The network topology refers to the spatial arrangements of its elements. The analysis of topological properties tells us how different nodes are connected to each other and how their communication paths look like. There are many aspects and characteristics that can be evaluated in a network. For simplicity we report just four metrics: number of nodes, number of edges, clustering coefficient and diameter.

The clustering coefficient is a measure of degree to which nodes in a network tend to cluster together. It expresses the likelihood that any two nodes with a common neighbour are themselves connected. The diameter indicates the maximum distance between two nodes in the network.

We compared the topology of the networks built with two different approaches: co-prediction and co-expression. For each generated network we calculated the topological properties described above. We report the comparison of FuNeL networks with co-expression networks inferred with different methods. The results considering Pearson are reported in Table S2, ARACNE is reported in Table S3 while MIC analysis is shown in Table S4.

![Table S2: Topological properties for co-prediction and Pearson co-expression networks generated for all 8 datasets.](image-url)
Table S3: Topological properties for co-prediction and ARACNE co-expression networks generated for all 8 datasets.

| Dataset  | Co-prediction | Co-expression (SE) | Co-expression (SN) |
|----------|---------------|--------------------|--------------------|
|          | Cat.          | $C_1$              | $C_2$              | $C_3$              | $C_4$              | $SE(C_1)$ | $SE(C_2)$ | $SE(C_3)$ | $SE(C_4)$ | $SN(C_1)$ | $SN(C_2)$ | $SN(C_3)$ | $SN(C_4)$ |
| Leukemia | Nodes         | 421                | 294                | 1480               | 988                | 9024      | 1426      | 1356      | 1577      | 422       | 1480      | 294       | 989       |
|          | Edges         | 1529               | 2154               | 2646               | 1529               | 2294      | 2154      | 2646      | 1529      | 512       | 2146      | 327       | 1479      |
|          | Clust.Coeff.  | 0.712              | 0.155              | 0.589              | 0.330              | 0.002     | 0.002     | 0.002     | 0.002     | 0.000     | 0.002     | 0.000     | 0.002     |
|          | Diameter      | 5                  | 6                  | 4                  | 6                  | 17        | 19        | 22        | 19        | 9         | 17        | 11        | 19        |
| LungH    | Nodes         | 429                | 382                | 1030               | 907                | 1614      | 1653      | 2066      | 429       | 1419      | 382       | 1030      |
|          | Edges         | 1068               | 2398               | 3410               | 1068               | 2317      | 2398      | 3410      | 435       | 1924      | 375       | 1250      |
|          | Clust.Coeff.  | 0.344              | 0.298              | 0.430              | 0.404              | 0.007     | 0.006     | 0.006     | 0.005     | 0.013     | 0.006     | 0.012     | 0.007     |
|          | Diameter      | 5                  | 8                  | 5                  | 7                  | 23        | 16        | 15        | 13        | 14        | 18        | 10        | 18        |
| LungM    | Nodes         | 91                 | 48                 | 247                | 143                | 1321      | 96        | 370       | 91        | 920       | 48        | 247       |
|          | Edges         | 134                | 78                 | 410                | 134                | 1858      | 78        | 410       | 72        | 1127      | 34        | 259       |
|          | Clust.Coeff.  | 0.379              | 0.262              | 0.418              | 0.475              | 0.000     | 0.002     | 0.000     | 0.009     | 0.000     | 0.005     | 0.000     | 0.014     |
|          | Diameter      | 3                  | 5                  | 3                  | 3                  | 13        | 17        | 11        | 18        | 11        | 17        | 5         | 11        |
| CNS      | Nodes         | 501                | 4257               | 494                | 3538               | 2092      | 4509      | 3584      | 5342      | 502       | 4257      | 494       | 3538      |
|          | Edges         | 4302               | 25039              | 12769              | 48940              | 4302      | 25069     | 12769     | 41661     | 513       | 20409     | 505       | 12458     |
|          | Clust.Coeff.  | 0.743              | 0.255              | 0.521              | 0.302              | 0.004     | 0.005     | 0.006     | 0.026     | 0.004     | 0.005     | 0.004     | 0.005     |
|          | Diameter      | 4                  | 7                  | 4                  | 6                  | 12        | 8         | 12        | 7         | 20        | 9         | 20        | 12        |
| Dlbcl    | Nodes         | 201                | 201                | 1617               | 380                | 1452      | 1191      | 2236      | 201       | 1699      | 201       | 1617      |
|          | Edges         | 848                | 10471              | 7351               | 3170               | 848       | 10471     | 7351      | 33890     | 269       | 14149     | 269       | 12903     |
|          | Clust.Coeff.  | 0.872              | 0.574              | 0.642              | 0.453              | 0.136     | 0.126     | 0.140     | 0.176     | 0.113     | 0.110     | 0.113     | 0.115     |
|          | Diameter      | 3                  | 5                  | 3                  | 5                  | 13        | 9         | 11        | 5         | 12        | 8         | 12        | 9         |
| GSE2191  | Nodes         | 890                | 846                | 3561               | 2574               | 5226      | 3846      | 5027      | 890       | 4802      | 846       | 3561      |
|          | Edges         | 3290               | 13424              | 6469               | 12074              | 3290      | 13424     | 6469      | 12076     | 846       | 10671     | 794       | 5564      |
|          | Clust.Coeff.  | 0.488              | 0.082              | 0.317              | 0.291              | 0.002     | 0.002     | 0.002     | 0.002     | 0.004     | 0.002     | 0.004     | 0.002     |
|          | Diameter      | 5                  | 9                  | 5                  | 9                  | 19        | 13        | 16        | 13        | 30        | 15        | 30        | 17        |
| GS3726   | Nodes         | 668                | 524                | 1170               | 1362               | 2167      | 1546      | 2279      | 668       | 2166      | 524       | 1170      |
|          | Edges         | 1761               | 2051               | 3502               | 1761               | 3255      | 2051      | 3502      | 787       | 3250      | 597       | 1455      |
|          | Clust.Coeff.  | 0.134              | 0.077              | 0.307              | 0.109              | 0.024     | 0.053     | 0.029     | 0.050     | 0.014     | 0.053     | 0.016     | 0.021     |
|          | Diameter      | 8                  | 10                 | 7                  | 8                  | 20        | 20        | 19        | 18        | 15        | 20        | 13        | 19        |
| Prostate | Nodes         | 938                | 704                | 2277               | 2760               | 6805      | 2268      | 4575      | 939       | 4290      | 704       | 2277      |
|          | Edges         | 3796               | 10175              | 3090               | 6546               | 3796      | 10175     | 3090      | 6546      | 1300      | 6095      | 1017      | 3192      |
|          | Clust.Coeff.  | 0.328              | 0.245              | 0.290              | 0.250              | 0.005     | 0.003     | 0.006     | 0.003     | 0.001     | 0.003     | 0.002     | 0.005     |
|          | Diameter      | 7                  | 10                 | 6                  | 8                  | 13        | 13        | 15        | 13        | 12        | 13        | 9         | 15        |
When analysing FuNeL networks we observed, as expected, that configurations having feature selection (C1 and C2) lead to networks with a smaller number of nodes than when the original set of attributes is used (C3 and C4). Furthermore, the second phase of machine learning modeling (C3 and C4) tends to reduce the number of nodes as it uses a reduced set of attributes as input (only significant nodes and their neighbours from the first training phase), while increasing both clustering coefficient and number of edges.

When comparing FuNeL and co-expression networks we notice that the ARACNE SE counterparts have in general more nodes. The same pattern can be found in SE(C2) and SE(C4) counterparts generated with Pearson and MIC, while it’s not true for the SE-networks based on configurations that use feature selection (C1 and C2). Conversely, SN-networks differ according to the inference method used. In fact ARACNE generated SN counterparts with less edges, while this is true only for SN(C1) and SN(C3) inferred with MIC and Pearson. The clustering coefficient is constantly lower in ARACNE networks than in FuNeL, this is probably due to the pruning phase operated by the method. A similar trend can be noticed for MIC networks with some exceptions (e.g Prostate SN(C2) and SN(C4)). A more balanced situation occurs when FuNeL is contrasted with Pearson, in fact networks generated with feature selection (C1 and C3) have a lower coefficient than their co-expression counterparts. Finally a clear pattern emerge when analysing the diameter of the networks. Co-prediction networks are always more compact than co-expression counterparts having up to 3 time lower diameter for MIC and Pearson and up to 7 time lower for ARACNE.
4 Enrichment score analysis

In this section we report the network average rankings, based on the Enrichment Score, across the 8 datasets for each inferring method. The networks are ranked between 1 and $N$ (where $N = 4$ for FuNeL and $N = 8$ for Pearson, ARACNE and MIC: $4 SE(C_1) + 4 SN(C_1)$). We considered Gene Ontology terms (biological process (BP), molecular function (MF) and cellular component (CC)) and biological pathways. The last row of each table represents the average rank across different biological categories.

Table S5: Average ranks based on the Enrichment Score across the 8 datasets for the networks generated with FuNeL.

| Cat.     | C1  | C2  | C3  | C4  |
|----------|-----|-----|-----|-----|
| GO BP    | 3   | 4   | 2   | 1   |
| GO MF    | 4   | 2.5 | 1   | 2.5 |
| GO CC    | 2   | 4   | 1   | 3   |
| Pathways | 4   | 2   | 3   | 1   |
| Average  | 3.25| 3.125| 1.75| 1.88|

Table S6: Average ranks based on the Enrichment Score across the 8 datasets for the networks generated with Pearson.

| Cat.     | Pearson (SE) | Pearson (SN) |
|----------|--------------|--------------|
|          | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 |
| GO BP    | 2  | 4  | 1  | 3  | 6  | 7  | 5  | 8  |
| GO MF    | 8  | 3.5| 5  | 6  | 7  | 3.5| 1  | 2  |
| GO CC    | 2  | 5  | 4  | 6  | 1  | 8  | 3  | 7  |
| Pathways | 6  | 5  | 4  | 3  | 7  | 1  | 8  | 2  |
| Average  | 4.5| 4.38| 3.5| 4.5| 5.25| 4.88| 4.25| 4.75|

Table S7: Average ranks based on the Enrichment Score across the 8 datasets for the networks generated with ARACNE.

| Cat.     | ARACNE (SE) | ARACNE (SN) |
|----------|-------------|-------------|
|          | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 |
| GO BP    | 4  | 7  | 5.5| 8  | 2  | 5.5| 1  | 3  |
| GO MF    | 4  | 8  | 4  | 7  | 2  | 6  | 1  | 4  |
| GO CC    | 3  | 7  | 5  | 8  | 2  | 6  | 1  | 4  |
| Pathways | 1  | 4  | 2  | 6  | 7  | 5  | 8  | 3  |
| Average  | 3  | 6.5| 4.13| 7.25| 3.25| 5.63| 2.75| 3.5|
Table S8: Average ranks based on the Enrichment Score across the 8 datasets for the networks generated with MIC.

| Cat.      | MIC (SE) | MIC (SN) |
|-----------|----------|----------|
|           | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
| GO BP     | 2     | 6     | 4     | 5     | 3     | 8     | 1     | 7     |
| GO MF     | 1     | 6     | 4     | 7     | 3     | 8     | 2     | 5     |
| GO CC     | 3     | 5,5   | 4     | 5,5   | 2     | 8     | 1     | 7     |
| Pathways  | 2,5   | 4     | 1     | 5,5   | 8     | 5,5   | 7     | 2,5   |
| Average   | 2.13  | 5.38  | 3.25  | 5.75  | 4     | 7.38  | 2.75  | 5.38  |

In here we also report the results of the analysis where we compared each inferring method against FuNeL using the Enrichment Score. The networks are ranked from 1 to 12: $4C_i + 4SE(C_i) + 4SN(C_i)$. In Table S9 we report, for each biological category and for each network, the ranks averaged across the 8 datasets. The row-wise rank is given in brackets and the highest ranks are shown with bold font. The following abbreviations were used for GO categories: biological process (BP), molecular function (MF) and cellular component (CC).

Table S9: Average network ranks based on the Enrichment Score.

| Method  | Co-prediction | Co-expression (SE) | Co-expression (SN) |
|---------|---------------|--------------------|--------------------|
|         | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $SE(C_1)$ | $SE(C_2)$ | $SE(C_3)$ | $SE(C_4)$ | $SN(C_1)$ | $SN(C_2)$ | $SN(C_3)$ | $SN(C_4)$ |
| Pearson | GO BP   | 6,6 (6) | 7,0 (7,5) | 7,0 (7,5) | 5,8 (3,5) | 5,8 (3,5) | 5,8 (3,5) | 5,12 (1) | 5,88 (1,5) | 7,06 (9) | 7,75 (12) | 7,12 (16) | 7,38 (11) |
|         | GO MF   | 7,8 (11) | 5,3 (2) | 6,19 (5) | 5,62 (3) | 9,12 (12) | 6,56 (7,5) | 6,50 (7,5) | 7,38 (10) | 7,19 (9) | 6,25 (6) | 4,31 (4) | 5,75 (4) |
|         | GO CC   | 4,31 (5) | 11,08 (12) | 4,19 (4) | 9,60 (10) | 3,88 (2) | 6,25 (6,5) | 6,25 (6,5) | 8,12 (8) | 3,19 (1) | 9,38 (11) | 4,06 (3) | 8,38 (9) |
|         | Pathways | 8,12 (10) | 4,75 (2) | 8,12 (10) | 4,38 (1) | 6,84 (8) | 6,50 (6) | 6,69 (7) | 5,88 (5) | 7,62 (9) | 5,00 (3) | 8,50 (12) | 5,50 (4) |
| ARACNE  | GO BP   | 6,6 (7) | 6,25 (5,5) | 6,94 (10) | 4,75 (1) | 6,25 (5,5) | 8,12 (11) | 6,88 (8,5) | 9,25 (12) | 5,19 (3) | 6,88 (8,5) | 5,06 (2) | 5,75 (4) |
|         | GO MF   | 7,44 (10) | 6,50 (8) | 6,19 (6,5) | 5,69 (3) | 6,00 (5) | 8,75 (12) | 5,62 (2) | 8,62 (11) | 5,81 (4) | 7,00 (9) | 4,19 (4) | 6,19 (6,5) |
|         | GO CC   | 4,31 (4) | 10,75 (12) | 3,44 (3) | 8,25 (6) | 5,50 (5) | 9,38 (10) | 6,75 (7) | 10,12 (11) | 2,44 (2) | 8,88 (9) | 2,31 (4) | 5,88 (6) |
|         | Pathways | 7,88 (10,5) | 5,38 (3) | 7,88 (10,5) | 5,25 (2) | 4,88 (1) | 6,25 (6) | 5,50 (4) | 7,00 (8) | 7,56 (9) | 6,50 (7) | 8,38 (12) | 5,50 (6) |
| MIC     | GO BP   | 7,44 (8,5) | 7,88 (11) | 7,44 (8,5) | 6,38 (5,5) | 4,12 (2) | 7,00 (7) | 6,60 (4) | 6,38 (5,5) | 4,81 (3) | 9,12 (12) | 3,94 (1) | 7,50 (10) |
|         | GO MF   | 8,06 (10) | 8,50 (12) | 7,19 (8) | 7,75 (9) | 3,62 (1) | 6,50 (5,5) | 5,12 (3) | 6,75 (7) | 5,31 (4) | 8,12 (11) | 4,56 (2) | 6,50 (5,5) |
|         | GO CC   | 5,19 (6) | 13,62 (12) | 4,44 (4) | 10,12 (10) | 4,25 (3) | 7,12 (7,5) | 5,12 (3) | 7,12 (7,5) | 3,19 (2) | 10,38 (11) | 1,69 (1) | 7,75 (9) |
|         | Pathways | 8,75 (12) | 4,75 (1) | 7,50 (10) | 5,58 (3) | 6,00 (5) | 6,50 (6) | 5,00 (2) | 6,02 (7) | 7,56 (11) | 7,00 (9) | 6,94 (8) | 5,88 (4) |

7
5 Disease association analysis

In this section we report the network average rankings across the 8 datasets for every inferring method based on the gene-disease association properties: participation in triangular relationship and proximity. We used two sources for the disease associations: Malacards [Rappaport et al., 2013] (a meta-database of human maladies consolidated from 64 independent sources) and manually curated databases (OMIM [Hamosh et al., 2002], Orphanet [Orphanet, 1997], Uniprot [Magrane and Consortium, 2011] and CTD [Davis et al., 2014]). The networks are ranked between 1 and \( N \) (where \( N = 4 \) for FuNeL and \( N = 8 \) for Pearson, ARACNE and MIC: \( 4 SE(C_i) + 4 SN(C_i) \)). The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A. The last row of each table represents the average rank across different metrics.

5.1 Malacards

Table S10: Average ranks based on the disease-association analysis (Malacards) across the 8 datasets for the networks generated with FuNeL.

|   | C1 | C2 | C3 | C4 |
|---|----|----|----|----|
| 1A | 3  | 1  | 4  | 2  |
| 2A | 4  | 1  | 2  | 3  |
| 3A | 3  | 3  | 1  | 3  |
| Proximity | 3  | 1  | 4  | 2  |
| Average | 3.25 | 1.5 | 2.75 | 2.5 |

Table S11: Average ranks based on the disease-association analysis (Malacards) across the 8 datasets for the networks generated with Pearson.

|   | Pearson (SE) | Pearson (SN) |
|---|--------------|--------------|
|   | \( C_1 \) | \( C_2 \) | \( C_3 \) | \( C_4 \) | \( C_1 \) | \( C_2 \) | \( C_3 \) | \( C_4 \) |
| 1A | 8  | 5  | 6  | 3  | 4  | 1  | 2  | 7  |
| 2A | 7  | 4  | 5  | 3  | 8  | 2  | 6  | 1  |
| 3A | 6.5 | 6.5 | 6.5 | 6.5 | 3  | 2  | 4  | 1  |
| Proximity | 1.5 | 5  | 3  | 1.5 | 7  | 6  | 8  | 4  |
| Average | 5.75 | 5.13 | 5.13 | 3.5 | 5.5 | 2.75 | 5  | 3.25 |

Table S12: Average ranks based on the disease-association analysis (Malacards) across the 8 datasets for the networks generated with ARACNE.

|   | ARACNE (SE) | ARACNE (SN) |
|---|------------|------------|
|   | \( C_1 \) | \( C_2 \) | \( C_3 \) | \( C_4 \) | \( C_1 \) | \( C_2 \) | \( C_3 \) | \( C_4 \) |
| 1A | 4  | 8  | 7  | 5.5 | 2.5 | 2.5 | 5.5 | 1  |
| 2A | 7  | 6  | 8  | 1  | 4  | 3  | 2  | 5  |
| 3A | 5.5 | 1  | 5.5 | 2  | 5.5 | 5.5 | 5.5 | 5.5 |
| Proximity | 7  | 3  | 5  | 1  | 6  | 2  | 8  | 4  |
| Average | 5.88 | 4.5 | 6.38 | 2.38 | 4.5 | 3.25 | 5.25 | 3.88 |
Table S13: Average ranks based on the disease-association analysis (Malacards) across the 8 datasets for the networks generated with MIC.

| Cat. | MIC (SE) | MIC (SN) |
|------|----------|----------|
|      | $C_1$  | $C_2$  | $C_3$  | $C_4$  | $C_1$  | $C_2$  | $C_3$  | $C_4$  |
| 1A   | 6      | 3      | 7.5    | 1      | 5      | 2      | 7.5    | 4      |
| 2A   | 8      | 3      | 5      | 2      | 7      | 1      | 6      | 4      |
| 3A   | 7      | 2      | 8      | 6      | 5      | 1      | 3      | 4      |
| Proximity | 6      | 1      | 2      | 5      | 7      | 4      | 8      | 3      |
| Average | 6.75  | 2.25  | 5.63   | 3.5    | 6      | 2      | 6.13   | 3.75   |

5.2 Curated databases

Table S14: Average ranks based on the disease-association analysis (curated databases) across the 8 datasets for the networks generated with FuNeL.

| Cat. | C1 | C2 | C3 | C4 |
|------|----|----|----|----|
| 1A   | 3  | 1  | 4  | 2  |
| 2A   | 3  | 4  | 1  | 2  |
| 3A   | 1  | 2.5| 2.5| 4  |
| Proximity | 4    | 1  | 3  | 2  |
| Average | 2.75 | 2.13 | 2.67 | 2.5 |

Table S15: Average ranks based on the disease-association analysis (curated databases) across the 8 datasets for the networks generated with Pearson.

| Cat. | Pearson (SE) | Pearson (SN) |
|------|--------------|--------------|
|      | $C_1$  | $C_2$  | $C_3$  | $C_4$  | $C_1$  | $C_2$  | $C_3$  | $C_4$  |
| 1A   | 8      | 4      | 6.5    | 1.5    | 6.5    | 1.5    | 3      | 5      |
| 2A   | 2      | 7      | 5.5    | 3      | 4      | 5.5    | 1      | 8      |
| 3A   | 5      | 7      | 4      | 8      | 1      | 6      | 3      | 2      |
| Proximity | 4     | 8      | 2.5    | 7      | 5      | 2.5    | 1      | 6      |
| Average | 4.5  | 6.5    | 4.63   | 4.88   | 5.13   | 3.88   | 2      | 5.25   |

Table S16: Average ranks based on the disease-association analysis (curated databases) across the 8 datasets for the networks generated with ARACNE.

| Cat. | ARACNE (SE) | ARACNE (SN) |
|------|-------------|-------------|
|      | $C_1$  | $C_2$  | $C_3$  | $C_4$  | $C_1$  | $C_2$  | $C_3$  | $C_4$  |
| 1A   | 3      | 5      | 8      | 6      | 2      | 4      | 1      | 7      |
| 2A   | 5      | 1      | 2      | 3      | 7.5    | 6      | 7.5    | 4      |
| 3A   | 4.5    | 4.5    | 4.5    | 4.5    | 4.5    | 4.5    | 4.5    | 4.5    |
| Proximity | 6     | 5      | 4      | 1      | 7.5    | 3      | 7.5    | 2      |
| Average | 4.63 | 3.88   | 4.63   | 3.63   | 5.38   | 4.38   | 5.13   | 4.38   |
Table S17: Average ranks based on the disease-association analysis (curated databases) across the 8 datasets for the networks generated with MIC.

| Cat.   | MIC (SE)    | MIC (SN)    |
|--------|-------------|-------------|
|        | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
| 1A     | 6     | 2     | 4     | 3     | 8     | 1     | 7     | 5     |
| 2A     | 7     | 3     | 8     | 1.5   | 5.5   | 1.5   | 5.5   | 4     |
| 3A     | 4     | 1     | 8     | 2     | 6     | 5     | 7     | 3     |
| Proximity | 5.5   | 1     | 3     | 4     | 7     | 2     | 8     | 5.5   |
| Average| 5.63  | 1.75  | 5.75  | 2.63  | 6.63  | 2.38  | 6.88  | 4.38  |
6 Case study: prostate cancer dataset

In this section we report the additional results from the analysis performed using the prostate dataset (Singh et al., 2002) as a case study. In particular we show: 1) the overlap of enriched terms between co-prediction and co-expression networks, 2) the overlap between GO terms associated to the hubs of the networks generated with different methods and FuNeL, and 3) the average percentages of alteration for key nodes of both co-prediction and co-expression networks in an independent dataset.

6.1 Overlap of networks enriched terms

We performed an analysis on the enriched terms of each network to highlight the complementary nature of the co-prediction and the co-expression paradigm. We generated heatmaps showing the unique terms associated only to co-prediction or co-expression networks. The main manuscript includes the comparison between FuNeL and Pearson networks, in here we report the analysis performed considering ARACNE (Figure S2) and MIC (Figure S3) networks. For the sake of readability we filtered out the generic GO terms (with depth < 9 in the GO hierarchical structure).

Figure S2: Number of non-common enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). On the x-axis we show the 12 investigated networks. On the y-axis we show the names of enriched terms unique to co-prediction or ARACNE co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

Figure S3: Number of non-common enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). On the x-axis we show the 12 investigated networks. On the y-axis we show the names of enriched terms unique to co-prediction or MIC co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.
When comparing ARACNE and FuNeL, we found 16 unique pathways for co-prediction networks and 8 for co-expression. In terms of unique GO terms, the overlap was more balanced, 7 for co-prediction networks and 9 for co-expression networks. \( C_2 \) and \( C_4 \), generated without feature selection, had the largest number of unique pathways, while \( SE(C_2) \) had the highest number of terms for ARACNE. The comparison of FuNeL with MIC generated many empty columns (Figure S3) for the GO terms because several networks resulted having no unique enriched terms. All the 15 unique GO terms related to MIC were associated to \( SN(C_2) \) (and with \( SN(C_4) \) in two cases), conversely FuNeL had more networks sharing the 12 unique terms. Finally, as noticed for in the ARACNE comparison, FuNeL networks are more enriched in biological pathways: 16 against 8 unique terms for MIC co-expression.

6.2 Overlap of hubs related terms

We also analysed the gene associated to the hubs of each network in order to compare the biological knowledge associated to them. A node \( v \) was considered to be a hub if its degree was at least one standard deviation above the mean network degree, that is if:

\[
d(v) > \mu_d + \sigma_d
\]

(4)

where \( d(v) \) is a degree of the node \( v \), and \( \mu_d \) and \( \sigma_d \) are the mean and standard deviation of a network node degree distribution.

To compare the networks, we used the 10 most frequent GO terms (biological processes) shared among each network’s hubs. Figure S4, S5 and S6 show the terms-overlap analysis between FuNeL networks and Pearson, ARACNE and MIC respectively. To make this analysis more specific we have discarded the most generic / most common terms (which could be be associated with many genes), we considered only the GO terms situated at level 10 of the GO hierarchy or lower. Blue terms were found only in co-prediction networks, red terms were found only in co-expression networks, and green terms were found in both. In Table S18 we summarise the number of unique and common terms shared between networks created with different approaches. This analysis further highlights the complementary nature of co-prediction and co-expression approach, the terms that are paradigm-specif always outnumber the common ones.

| Terms        | Pearson | ARACNE | MIC |
|--------------|---------|--------|-----|
| Co-prediction| 16      | 18     | 16  |
| Co-expression| 19      | 20     | 19  |
| Common       | 11      | 9      | 11  |

Table S18: Unique and common terms from networks’ hubs
**Figure S4**: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and Pearson co-expression networks.
### Figure S5: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and ARACNE co-expression networks.

| Process                                                                 | GO ID       |
|------------------------------------------------------------------------|-------------|
| positive regulation of glucocorticoid receptor signaling pathway       | GO:2000324  |
| negative regulation of BMP signaling pathway                            | GO:0030514  |
| positive regulation of peptidyl-tyrosine phosphorylation               | GO:0033138  |
| toll-like receptor TLR6:TLR2 signaling pathway                          | GO:0038124  |
| signal transduction involved in DNA damage checkpoint                  | GO:0072422  |
| positive regulation of protein autophosphorylation                      | GO:0031954  |
| regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion | GO:0014009  |
| activation of phosphatase C activity                                    | GO:0007202  |
| negative regulation of activation-induced cell death of T cells        | GO:0070236  |
| negative regulation of histone deacetylation                           | GO:0031064  |
| positive regulation of peptidyl-lysine phosphorylation                 | GO:0050731  |
| positive regulation of BRK1 and ERK2 cascade                           | GO:0070374  |
| negative regulation of neuron projection development                    | GO:0010977  |
| activation of protein kinase A activity                                 | GO:0034199  |
| transforming growth factor beta receptor complex assembly               | GO:0000771  |
| negative regulation of BRK1 and ERK2 cascade                           | GO:0070373  |
| regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion | GO:0010981  |
| activation of MAPK activity                                             | GO:0000187  |
| MyD88-dependent toll-like receptor signaling pathway                   | GO:002759   |
| MyD88-independent toll-like receptor signaling pathway                 | GO:002759   |
| cellular iron ion homeostasis                                          | GO:0006879  |
| positive regulation of adenylate cyclase activity involved in G protein coupled receptor signaling pathway | GO:0020259 |
| positive regulation of activin receptor signaling pathway              | GO:0032927  |
| protein K11-linked ubiquitination                                      | GO:0070979  |
| positive regulation of sodium ion export from cell                     | GO:1903278  |
| positive regulation of CMAP biosynthetic process                       | GO:0030819  |
| activation of protein kinase A activity                                 | GO:0032147  |
| anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process | GO:0031145 |
| calcium ion transport                                                  | GO:0006816  |
| toll-like receptor 9 signaling pathway                                 | GO:0034162  |
| toll-like receptor 5 signaling pathway                                 | GO:0034146  |
| negative regulation of telomere single strand break repair            | GO:1903824  |
| toll-like receptor 4 signaling pathway                                 | GO:0034142  |
| DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest | GO:0006977  |
| positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition | GO:006437   |
| sodium ion export from cell                                            | GO:0036376  |
| protein polyubiquitination                                             | GO:0000020  |
| activation of MAPKK activity                                           | GO:0000186  |
| toll-like receptor 3 signaling pathway                                 | GO:0034138  |
| toll-like receptor 2 signaling pathway                                 | GO:0034154  |
| positive regulation of cytosolic calcium ion concentration            | GO:0007204  |
| toll-like receptor TLR1:TLR2 signaling pathway                         | GO:0038123  |
| epidermal growth factor receptor signaling pathway                    | GO:0007173  |
| positive regulation of neuron projection development                   | GO:0010976  |
| TRIF-dependent toll-like receptor signaling pathway                   | GO:0034566  |
| stimulatory C-type lectin receptor signaling pathway                   | GO:0002213  |
Figure S6: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and MIC co-expression networks.
6.3 Validation on independent dataset

In this section we report additional informations about the analysis performed using data from the independent prostate cancer study (Taylor et al., 2010) available in the cBioPortal for Cancer Genomics (Cerami et al., 2012). In particular we report the full list of alterations for the topologically important genes analysed in the main article. The Figure S7–S14 show the percentage of altered tumour samples for top 10 hubs (nodes with highest degree) and top 10 central nodes (with highest betweenness centrality) in the best performing networks according to the gene-disease association analysis (using the information from the curated databases). The selected networks are $C_2$ for FuNeL, $SN(C_3)$ for Pearson, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC. For all of them we report the alterations for both hubs and central nodes.

| Genetic Alteration | PTGDS | PAGE4 | LMO3 | GSTM2 | NELL2 | COL4A6 | MAF | ABL1 | RBP1 | PARM1 | MYH11 |
|--------------------|-------|-------|------|-------|-------|--------|-----|------|------|-------|-------|
| Deep Deletion      | 45%   | 28%   | 11%  | 52%   | 9%    | 65%    | 24% | 11%  | 20%  | 53%   | 58%   |
| Missense Mutation  |       |       |      |       |       |        |     |      |      |       |       |
| mRNA Upregulation  |       |       |      |       |       |        |     |      |      |       |       |
| mRNA Downregulation|       |       |      |       |       |        |     |      |      |       |       |

**Figure S7:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest degree (hubs) in $C_2$ network are shown.

| Genetic Alteration | PTGDS | PAGE4 | GSTM2 | NELL2 | COL4A6 | MAF | ABL1 | RBP1 | PARM1 | MYH11 |
|--------------------|-------|-------|-------|-------|--------|-----|------|------|-------|-------|
| Deep Deletion      | 45%   | 28%   | 52%   | 9%    | 65%    | 24% | 11%  | 20%  | 53%   | 58%   |
| Missense Mutation  |       |       |       |       |        |     |      |      |       |       |
| mRNA Upregulation  |       |       |       |       |        |     |      |      |       |       |
| mRNA Downregulation|       |       |       |       |        |     |      |      |       |       |

**Figure S8:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in $C_2$ network are shown.
**Figure S9:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest degree** (hubs) in Pearson $SN(C_3)$ network are shown.

**Figure S10:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest betweenness centrality** (central nodes) in Pearson $SN(C_3)$ network are shown.

**Figure S11:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest degree** (hubs) in ARACNE $SE(C_4)$ network are shown.
| Gene     | Percentage |
|----------|------------|
| IL10RA   | 16%        |
| THRA     | 33%        |
| PSG1     | 16%        |
| SAMD14   | 1%         |
| BTF3     | 13%        |
| GNAS     | 13%        |
| ALAS2    | 4%         |
| RHBDL1   | 8%         |
| MYBPC3   | 4%         |
| DDX11    | 14%        |

**Figure S12:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in ARACNE $SE(C_2)$ network are shown.

| Gene     | Percentage |
|----------|------------|
| WBSCR22  | 39%        |
| ADAM15   | 12%        |
| GP2      | 7%         |
| COP56    | 14%        |
| RIMS2    | 16%        |
| ATP1B2   | 18%        |
| CEP170B  | 21%        |
| POU4F1   | 1%         |
| HAUS5    | 19%        |
| PCBP3    | 5%         |

**Figure S13:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest degree (hubs) in MIC $SE(C_2)$ network are shown.

| Gene     | Percentage |
|----------|------------|
| ADAM15   | 12%        |
| WBSCR22  | 39%        |
| COP56    | 14%        |
| CEP170B  | 21%        |
| CPNE6    | 34%        |
| GP2      | 7%         |
| ATP1B2   | 18%        |
| EFNA3    | 9%         |
| UBE3B    | 13%        |
| PAX8     | 4%         |

**Figure S14:** Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with highest betweenness centrality (central nodes) in MIC $SE(C_2)$ network are shown.
7 Visualisation of the co-prediction and co-expression networks

In this section we include the layouts of co-prediction and co-expression networks for three of the datasets used in the main article: Prostate (used in the case study) and Lung-Michigan (small networks). Only a few examples are shown, for which the differences in the topology of the networks generated with the two approaches is the most visible. The networks were visualised using the Organic Layout in Cytoscape [Shannon et al., 2003].

Prostate: FuNeL – $C_1$  
Prostate: FuNeL – $C_3$

Prostate: ARACNE – $SN(C_1)$  
Prostate: ARACNE – $SE(C_3)$
Lung-Michigan: FuNeL – $C_2$

Lung-Michigan: ARACNE – $SE(C_2)$

Lung-Michigan: FuNeL – $C_4$

Lung-Michigan: ARACNE – $SN(C_4)$
References

Cerami, E. et al (2012). The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery, 2*(5), 401–404.

Chowdary, D. et al (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The Journal of Molecular Diagnostics, 8*(1), 31 – 39.

Davis, A.P. et al (2014). The comparative toxicogenomics database’s 10th year anniversary: update 2015. *Nucleic Acids Research*.

Hamosh, A. et al (2002). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, pages 52–55.

Magrane, M. et al (2011). Uniprot knowledgebase: a hub of integrated protein data. *Database, 2011*.

Orphanet (1997). Orphanet: an online database of rare diseases and orphan drugs. copyright, inserm 1997.

Rappaport, N. et al (2013). Malacards: an integrated compendium for diseases and their annotation. *Database, 2013*.

Shannon, P. et al (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498–2504.

Singh, D. et al (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1*(2), 203 – 209.

Taylor, B.S. et al (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell, 18*(1), 11 – 22.