Supplementary Information for ONCOPLEX: An Oncology-Inspired Hypergraph Model Integrating Diverse Biological Knowledge for Cancer Driver Gene Prediction

1 Experimental setting

1.1 Hyperparameters optimization

A successful machine learning model depends on selecting suitable hyperparameters, making careful tuning essential. We performed hyperparameter optimization using 5-fold cross validation. The labeled data were divided into K folds, where K-1 folds were used for training and one fold was held out for testing. Within each training fold, an inner loop applied a simple grid search to explore combinations of hyperparameters and identify the optimal configuration based on validation performance. We tuned the following hyperparameters:

- Number of hidden units: `[64, 128, 256]`
- Number of layers: `[2, 3, 4]`
- Learning rate: `[1e-3, 5e-4]`
- Weight decay: `[1e-4, 1e-3]`
- Dropout: `[0.4, 0.5, 0.25]`
- Positive class weight: `[0.4, 0.2, 0.45]`

The **outer loop** was then used to evaluate the model with the best hyperparameter combination, and report the average and slandered deviation for the three metrics AUPRC, AUROC, and the F1 score. This evaluation provide an unbiased and robust estimate of the model performance. We repeat the same procedures for the specific cancer setting. In some cancer types, the best set of hyperparameters is identical, while in other cases, a different set is optimal for that specific cancer.

1.2 Datasets

1.2.1 Driver genes labels

We collected the driver gene sets for both pan-cancer and cancer-specific driver genes from the following resources:

- Network of Cancer Genes (NCG) v6.0 [8]:
  `http://ncg.kcl.ac.uk/download_file.php?file=cancergenes_list.txt`
- DigSee Database [3]:
  `http://210.107.182.61/digseeOld/`
- COSMIC Cancer Gene Census (CGC) v91 and COSMIC Mutations in Census Genes[10]:
  `https://cancer.sanger.ac.uk/cosmic/download`
- IntOGen tumor-specific driver lists v2024.09.20 [7] `https://www.intogen.org/download`

1.2.2 Gene sets used for biological validation

To further confirm our results, we assessed whether the newly predicted genes appeared in the following cancer gene databases:

- **OncoKB**: A manually curated resource of cancer genes annotated based on validated oncogenic effects. Only high-confidence cancer genes—those supported by evidence from more than three independent sources, including clinical studies and association with FDA-approved drugs—were included in this set [1].
  `https://www.oncokb.org/cancerGenes`
- **ONGene**: A literature-curated collection of human oncogenes [5].
  `http://ongene.bioinfo-minzhao.org/ongene_human.txt`
- **CancerMine**: A text-mining–based database of cancer genes extracted from published literature [4].

2 Results

2.1 Analysis of newly predicted genes by OncoPlex across 11 cancer types

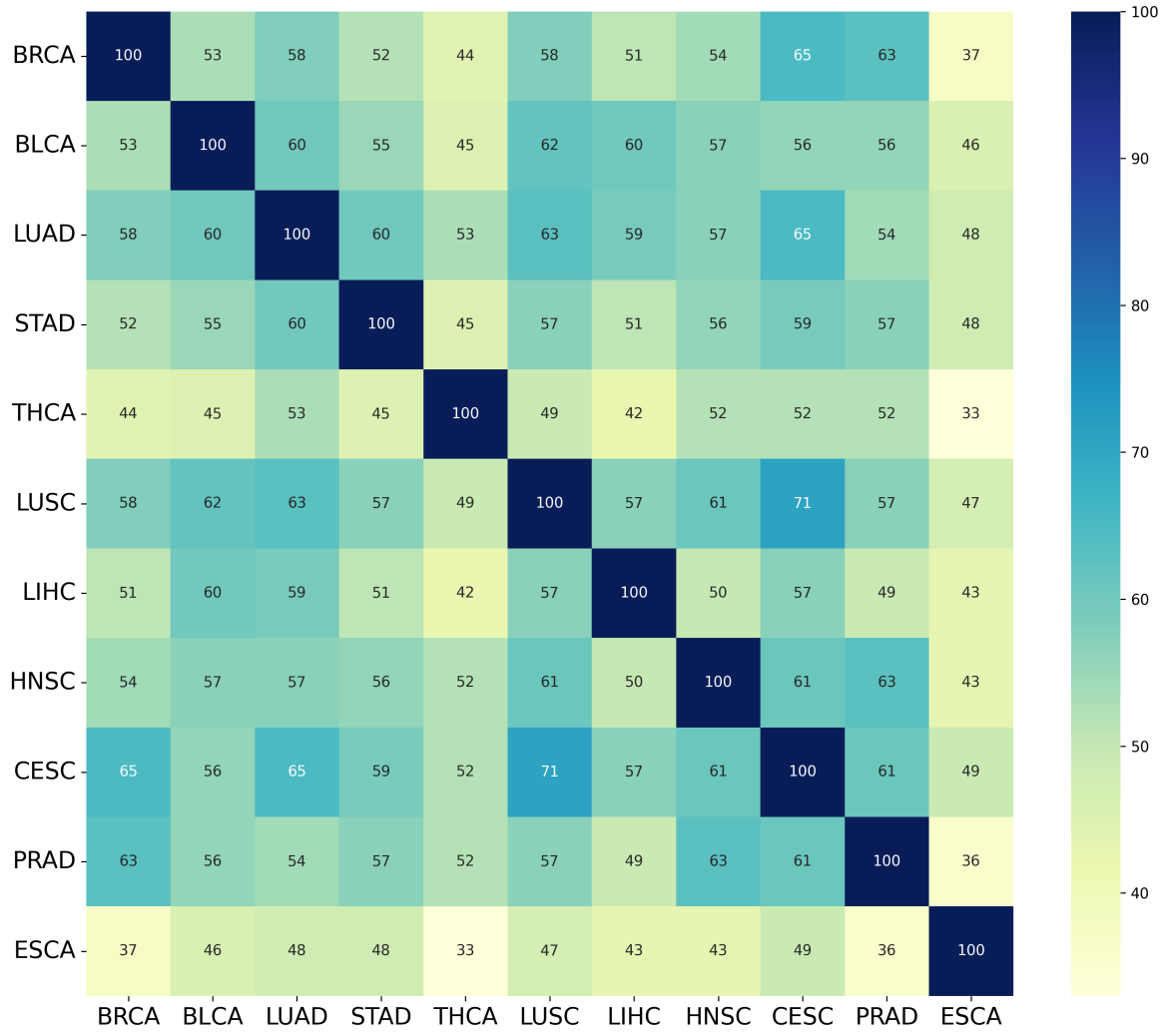

Figure 1: The overlap among the 100 newly predicted genes across different cancer types. For each cancer type, we evaluate how its predicted genes intersect with those of other cancers. As shown, some cancer types—such as BRCA—exhibit high levels of overlap with many others, suggesting shared molecular characteristics. In contrast, cancers like ESCA show minimal overlap with most other types, except for specific cases such as CESC, STAD, and LUAD, where moderate gene-level similarities are observed.

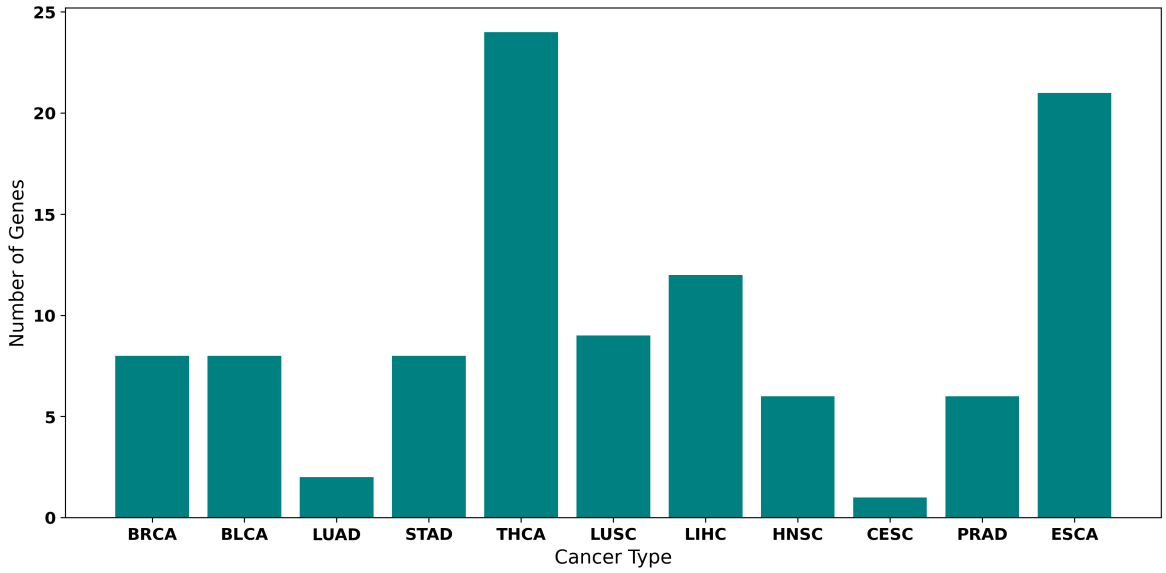

Figure 2: The number of unique predicted genes identified for each cancer type. We analyze the genes that are exclusively predicted for a specific cancer type. Among all cancers, THCA and ESCA exhibit the highest number of unique genes, which can be attributed to their limited overlap in pathways with other cancers. This reduced pathway overlap results in fewer shared genes with other cancer types.

3 Methods

3.1 Node features

We used three types of node features: **Incidence-based features**, **core omics features**, and **comprehensive features**.

Incidence-based features. These features were derived solely from the hypergraph topology, without relying on any predefined biological node features. Each gene is represented by an binary matrix that captures its position within the pathway hypergraph. This setup ensure that the only source of information comes from the higher order pathways connection.

Core omics features. These features were derived from multi-omics data collected across 16 cancer types in the TCGA database [2]. For each cancer type, we included three omics layers: Single-Nucleotide Variants (SNVs), gene expression, and DNA methylation. The preprocessing steps following [9] are detailed below:

- **Single nucleotide variants:** We extracted the single point point variant from TCGA Mutation Annotation Format (MAF) files. We computed the mutation

frequency mf for each gene by normalizing the number of non-silent mutations by gene length:

$$mf = \frac{\sum m}{L} \tag{1}$$

where m is the number of mutations and L is the gene length. We excluded ultra-mutated samples to reduce noise due to genome instability.

- **Gene expression:** TCGA provided gene expression data as normalized FPKM values. After batch correction, we computed the $\log_2$ fold change between tumor and matched normal samples:

$$\log_2 FC(i,j) = \log_2 \left(\frac{i}{j} \right) \tag{2}$$

- **DNA methylation:** We used beta values (ranging from 0 to 1) to compute differentially methylated regions. The $\log_2$ fold change at the gene level was calculated as:

$$dmc_i = \frac{1}{|S_c|} \sum_{s \in S_c} \log_2 \left(\frac{\beta_i^t}{\beta_i^n} \right) \tag{3}$$

where β_i^t and β_i^n are the beta values for tumor and normal samples, respectively.

Each omics type generated a gene-by-sample matrix, and we combined the three matrices per cancer type to form a final multi-dimensional feature representation. In the pan-cancer setting, this resulted in a **48** dimensional feature vector per gene. For cancer-type-specific networks, we used only the corresponding omics features from that cancer, resulting in a **3** dimensional feature vector per gene.

Comprehensive features. The last set of node features, referred to as comprehensive features, is a 44 genomics, epigenetics, and functional features collected from previous cancer driver prediction methods[6]. These features include gene-level biological properties such as histone modifications, evolutionary conservation, and gain/loss of function indicators. They are particularly useful in distinguishing oncogenes from tumor suppressor genes.

In the pan-cancer graph, we trained using all **44 comprehensive features**. In cancer-specific settings, we integrated these comprehensive features with the cancer-specific omics features described above, enabling the model to jointly learn both cancer-type-specific and shared cross-cancer properties. (See Supplementary Table 1).

Supplementary Tables

Supplementary Table 1: Comprehensive features

In this table, we present the details of the comprehensive features used for training in the second experiment we did, which are employed to distinguish between tumor suppressors and oncogene drivers across cancer types [6]. The 44 features include various types of mutations, such as missense and loss-of-function mutations, as well as epigenetic signals like DNA methylation and histone modifications. Additionally, a phenotype-based feature is included to enhance predictive performance.

Supplementary Table 2: Specific cancer driver and passenger genes statistics

This table shows the number of driver genes set for each cancer type, based on COSMIC and DigSEE databases. The passenger genes are the same across all cancer types, consistent with the pan-cancer setting described previously in the main article. The number of genes reported here indicate those included in our hypergraph after construction. All genes are actually present in our network without any missing gene.

Supplementary Table 3: Pan cancer results

In this table, we present the complete results for OncoPlex and all baseline methods on the pan-cancer dataset. The metrics include AUPRC, AUROC, and F1 score, each reported with its standard deviation.

Supplementary Table 4: Pan cancer ranking results

In this table, we present the ranking metrics—including precision and cumulative hit counts (Hits)— for values of k ranging from 1 to 50, evaluated for OncoPlex and all baseline methods. We choose k up to 50 because the number of known pan-cancer driver genes is relatively high, and larger k values help assess the models' ability to identify and rank a broader set of true positives among these drivers.

Supplementary Table 5: Cancer specific results

Here, we present the complete results for OncoPlex and all baseline methods on the cancer specific dataset. The metrics include AUPRC, AUROC, and F1 score, each reported with its standard deviation.

Supplementary Table 6: Cancer specific Precision@k results

In this table, we present the precision at k values of 1, 3, 5, and 10, evaluated for OncoPlex and all baseline methods on cancer-specific datasets. Unlike the pan-cancer setting, where the number of driver genes is large, the number of known driver genes for individual cancer types is relatively small. Therefore, we select smaller k values, as the maximum number of driver genes for a single cancer type—in BRCA—is less than 50.

Supplementary Table 6: Cancer specific Hits@k results

In this table, similar to the precision table above, we report the raw count of recovered true driver genes at k values of 1, 3, 5, and 10.

Supplementary Table 8: New predicted genes

In this table, we present the complete list of 30 newly candidate genes for each cancer type, classifying them as either common drivers or entirely novel genes that have not been previously reported as cancer driver genes. Additionally, we include supporting evidence from well-known databases. Most of the newly identified genes are supported by at least one database as being cancer-related, while a few rare genes have not been reported in any updated cancer-related databases. We refer to these genes as the novel driver list and present them in the following table.

Supplementary Table 9: Novel cancer genes

In this table, We present results for selected cancer types in which novel genes appeared among the top-ranking predictions. In other cancer types, no novel genes were observed in the top predictions. Then, to explore their potential significance, we conducted a comprehensive PubMed review to examine their functions and pathways, focusing on their possible roles in cancer. These genes might serve as biomarkers for further study and validation.

Supplementary Table 10: KEGG pathway enrichment analysis for the NPGs

In this table, we show the KEGG pathway enrichment analysis for the newly predicted genes in BRCA, HNSC, and STAD cancers. Only the pathways with significant p-value are selected.

Supplementary Table 11: Hallmark enrichment analysis for the NPGs

Similar to the pathways table, here we present the significant hallmarks for the new predicted genes in each cancer.

References

[1] Debyani Chakravarty et al. "OncoKB: A Precision Oncology Knowledge Base". In: *JCO Precis Oncol* (May 2017), pp. 1–16. DOI: https://doi.org/10.1200/PO.17.00011.

[2] Kyle Chang et al. "The Cancer Genome Atlas Pan-Cancer Analysis Project". In: *Nature Genetics* 45 (Sept. 2013), pp. 1113–1120. DOI: https://doi.org/10.1038/ng.2764.

[3] Jeongkyun Kim et al. "DigSee: Disease gene search engine with evidence sentences (version cancer)". In: *Nucleic acids research* 41 (June 2013). DOI: https://doi.org/10.1093/nar/gkt531.

[4] Jake Lever et al. "CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer". In: *Nat Methods* 16.6 (2019). DOI: https://doi.org/10.1038/s41592-019-0422-y.

[5] Yining Liu, Jingchun Sun, and Min Zhao. "ONGene: A literature-based database for human oncogenes". In: *Journal of Genetics and Genomics* 44 (2017), pp. 119–121. DOI: https://doi.org/10.1016/j.jgg.2016.12.004.

[6] Jie Lyu et al. "DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features". In: *Science Advances* 6 (Nov. 2020), eaba6784. DOI: https://doi.org/10.1126/sciadv.aba6784.

[7] Francisco Martínez-Jiménez et al. "A compendium of mutational cancer driver genes". In: *Nature Reviews Cancer* 20 (Oct. 2020). ISSN: 1474-1768. DOI: https://doi.org/10.1038/s41568-020-0290-x.

[8] Dimitra Repana et al. "The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens". In: *Genome Biology* 20 (Jan. 2019). DOI: https://doi.org/10.1186/s13059-018-1612-0.

[9] Roman Schulte-Sasse et al. "Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms". In: *Nature Machine Intelligence* 3 (June 2021), pp. 1–14. DOI: 10.1038/s42256-021-00325-y.

[10] Zbyslaw Sondka et al. "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers". In: *Nature Reviews Cancer* 18 (Oct. 2018). DOI: https://doi.org/10.1038/s41568-018-0060-1.