HELP: A computational framework for labelling and predicting human context-specific essential genes

Ilaria Granata[1*], Lucia Maddalena[1], Mario Manzo[2], Mario Rosario Guarracino[3], Maurizio Giordano[1]

1 Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy
2 Information Technology Services, University of Naples "L'Orientale", Naples, Italy
3 Department of Economics and Law, University of Cassino and Southern Lazio, Cassino, Frosinone, Italy

* ilaria.granata@cnr.it

Abstract

Machine learning-based approaches are particularly suitable for identifying essential genes as they allow the generation of predictive models trained on features from multi-source data. Gene essentiality is neither binary nor static but determined by the context. The databases for essential gene annotation do not permit the personalisation of the context, and their update can be slower than the publication of new experimental data. We propose HELP (Human Gene Essentiality Labelling & Prediction), a computational framework for labelling and predicting essential genes. Its double scope allows for identifying genes based on dependency or not on experimental data. The effectiveness of the labelling method was demonstrated by comparing it with other approaches in overlapping the state-of-the-art EG annotations, where HELP demonstrated the best compromise between false and true positive rates. The gene attributes, including multi-omics and network embedding features, lead to high-performance prediction of EGs while confirming the existence of essentiality nuances.

Author summary

Essential genes (EGs) are commonly defined as those required for an organism or cell's growth and survival. The essentiality is strictly dependent on both environmental and genetic conditions, determining a difference between those considered common EGs (cEGs), essential in most of the contexts considered, and those essential specifically to one or few contexts (context-specific EGs, csEGs). In this paper, we present a library of tools to address the identification and prediction of csEGs. Furthermore, we attempt to experimentally explore the statement that essentiality is not a binary property by identifying, predicting and analyzing an intermediate class between the Essential (E) and Not Essential (NE) genes. Among the multi-source data used to predict the EGs, we found the best attributes combination to capture the essentiality. We demonstrated that the additional class of genes we defined as "almost Essential" shows differences in these attributes from the E and NE genes. We believe that investigating the context-specificity and the dynamism of essentiality is particularly relevant to unravelling crucial insights into biological mechanisms and suggesting new candidates for precision medicine.

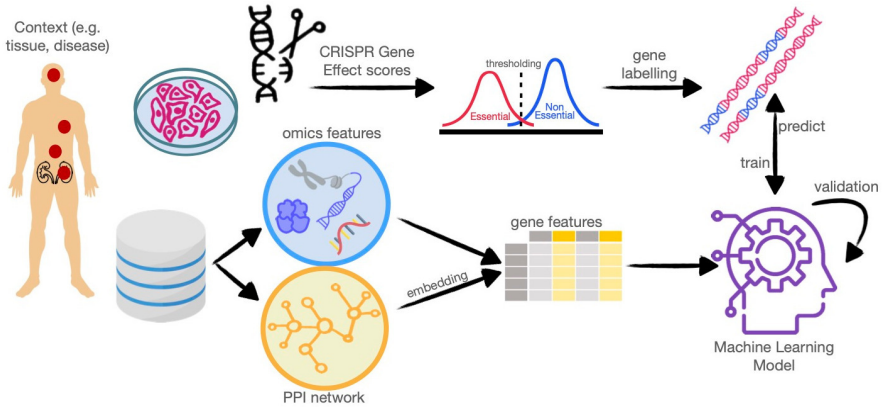


Fig 1. Schematic illustration of the HELP framework. HELP can be applied to each desired context (e.g. tissue, disease). The context guides the selection of gene knockout experiments and the collection of omics data and PPI. The CRISPR effect scores are used to derive the essentiality labels through an unsupervised thresholding approach. The omics and PPI embedding features are the input to train the machine learning prediction model.

Introduction

Identifying essential genes (EGs) is challenging and involves multiple disciplines and research areas. EGs are generally defined as necessary for the growth and survival of any organism or cell. Still, the definition is critical, as the term "essential" requires a specific contextualisation. In this scenario, a crucial role is played by the conditions in which the experimental procedures for the EGs recognition are performed. EGs are commonly identified through *in vitro* experiments on cell lines aimed at deleting the gene of interest and observing the effects on the phenotype. Single gene deletion, antisense RNA, transposon mutagenesis and CRISPR-Cas9 are the most used techniques. The latter is considered the state-of-the-art method for simplicity and efficiency [1]. At a genome-wide level, these procedures must be performed massively, becoming complex, costly, labour- and time-intensive. Computational methods to support the experimental approaches, minimising the costs and overcoming limitations related to the availability of *in vitro* models, are urgently needed. Approaches based on Machine Learning (ML) allow the generation of predictive models based on features coming from multi-modal, multi-source and multi-omics data, an aspect of primary importance in the era of precision medicine and the need for procedures able to capture context-specificity. The gene essentiality prediction is usually treated as a supervised classification problem, where the model is trained by using several characteristics of genes that are *a priori* labelled as Essential (E) or Not Essential (NE) [2,3]. An efficient model requires both consistent data and labels. Along with the biological and genetic features that can capture the gene essentiality, a set of information can be derived from networks representing the interactions of biological factors. The information contained in these systems can be learnt unsupervised through Deep Learning (DL) techniques [4–6]. The network primarily used in the context of EGs prediction is the Protein-Protein Interaction (PPI) network, describing the physical connections among proteins. According to the centrality-lethality rule, the more central a gene, or its product, the higher its probability of being essential [7]. While the massive experiments on multiple cell lines and conditions evidenced that the essentiality of a gene is strongly dependent on the context and that gene essentiality is neither a binary nor a static

property [8,9], the computational approaches proposed seem to ignore these fundamental aspects, relying on a binary classification of organism-wide EGs.

A crucial issue concerning the prediction of EGs is the labelling process. In most cases, the annotation of human E genes is retrieved from dedicated databases. An important resource is the Online GEne Essentiality (OGEE) database [10], which provides both organism- and tissue-level data and the relative lists of EGs. The labels derive from gene knockout experiments that produce scores reflecting cell fitness in the wake of the deletion of a specific gene. However, a pre-compiled list represents a limitation for context-specificity where the context of interest can include or exclude some experiments. Furthermore, the update of these databases is often slower than the publishing of new experimental data. Beyond databases, some tools for identifying EGs based on gene deletion scores have been developed [11].

To address the aspects mentioned above, we present HELP (Human Gene Essentiality Labelling & Prediction), a computational framework for context-specific EGs prediction that treats both the labelling and classification tasks. HELP, schematically described in Fig 1, computes the labelling of genes as E/NE through an unsupervised approach. We collected multi-source genetic features and the embedding extracted from a context-specific PPI network by a DL approach to investigate what characterises and can predict EGs. We individuated the best combination of gene attributes to capture the essentiality and demonstrated that integrating omics and network features improves the prediction performance. Finally, we exploited this flexible labelling method to explore the existence of essentiality shades.

Materials and methods

Labelling

The HELP framework labels genes based on the scores derived from gene knockout experiments. In particular, in the current work, we used the scores reported in the Gene Effect file (DepMap v. 23Q4; https://depmap.org/portal), derived from the CRISPR knockout screens published by Broad's Achilles and Sanger's SCORE projects. As described in DepMap, negative scores imply inhibition of cell growth and/or death following gene knockout. The assignment of genes to the two classes, E and NE, is obtained by thresholding their CRISPR scores for each cell line. The thresholds are determined by minimising intra-class intensity variance, or equivalently, by maximising inter-class variance, based on the Otsu method [12]. Through the labelling step, HELP identifies common EGs (cEGs), context-specific EGs (csEGs) and uncommon context-specific EGs (ucsEGs), where the context is a user-defined parameter. Here, we focused on two contexts corresponding to two human tissues: kidney and lung.

In the case of cEGs, we filtered the CRISPR data by discarding tissues with less than ten cell lines and genes with more than 95% missing values across all cell lines. HELP calculates the labels for each cell line separately, obtaining as many labels as the number of cell lines, and then computes tissue-specific gene labels as the mode of the labels of cell lines belonging to the same tissue. Finally, the mode of the modes gives the ultimate label of cEGs. This strategy avoids bias due to the different number of cell lines per tissue. In the case of ambiguities, the label is assigned to the NE class.

The labels of csEGs are computed for the context of interest in the same manner explained above, but excluding the calculation of the mode of the modes, specific to the cEGs. Also in this case, we adopted a gene filtering step excluding those with more than 95% missing values.

The ucsEGs are obtained by subtracting cEGs from the csEGs.

Besides the two usually considered classes E/NE, leading to what will be referred to

as "two-class HELP labelling", we aimed to investigate the existence of an intermediate class of context-specific genes. To this end, in the case of csEGs, we subdivided the class NE into two further classes, named aE (almost Essential) and sNE (strongly Not Essential), using the same procedure described for the two-class HELP labelling. In case of ambiguities, the label was assigned to the sNE class. This labelling will be referred to as the "three-class HELP labelling".

Prediction

Data: PPI embedding features

The human PPI network was downloaded from the Integrated Interaction Database (IID) [13]. The edges of the PPI are enriched with several annotations, including the tissue for which the interaction is reported and/or predicted. Filtering the PPI by tissue, we obtained kidney- and lung-specific PPI networks of 19314 nodes/1110251 edges and 19334 nodes/1111550 edges, respectively. Topological attributes of the PPIs were extracted using the node2vec algorithm [14]. It learns *node embeddings*, i.e. compact but informative vector-based representations of network nodes, by using the network topology as learning paths and maximising a neighbourhood-preserving objective function. The approach is flexible because the core of the embedding mechanism is a skipgram neural model [15] trained on simulated random walks and biased to allow different definitions of node neighbourhoods.

Data: Multi-omics features

To achieve the prediction task and investigate the characteristics descriptive of essentiality, we collected generic and tissue-specific genetic attributes from several sources. We generated 4710 (kidney) and 4707 (lung) numeric features regarding genomic, transcriptomic, epigenetic, functional and evolutionary characteristics for the two tissues under study, summarised and detailed in Table S1.

Structural information was represented by attributes such as "Gene length", "GC content", and "Transcript counts".

Four gene expression-related attributes were considered: "GTEX_*" (*= Kidney or Lung), mean of median-normalised gene expression values in kidney cortex and medulla or lung from GTEX portal [16]; "UP_tissue", the number of tissues in which the query gene has been found expressed as annotated in DAVID [17]; "OncoDB_expression", mean of the log_2Fold-Change values as results of the differential expression analysis of three renal cancer (KIRC, KIRP, KIRC) or two lung cancer (LUAD, LUSC) molecular subtypes vs normal samples [18]. Only genes with adjusted p-value $\leq$0.05 were considered; "HPA_*" (*= Kidney or Lung), normalised transcript expression values summarised by genes from RNA-sequencing experiments [19].

Functional characteristics were retrieved from Gene Ontology (GO), KEGG and REACTOME annotations. To convert the textual annotations into numerical attributes, for each query gene, we calculated the number of GO-Molecular Functions ("GO-MF"), GO-Biological Processes ("GO-BP"), GO-Cellular Components ("GO-CC"), KEGG ("KEGG") and REACTOME ("REACTOME") annotations. Attributes concerning protein sub-cellular localisation were also collected as confidence scores assigned to each of the 3305 GO-CC terms retrieved ("CCcfs") from COMPARTMENTS [20].

As EGs are genes strongly involved in physical and functional interactions, we considered four interactions-based attributes: "BIOGRID", the number of interactions annotated for each query gene [21]; "UCSC_TFBS", the count of predicted transcription factors binding sites (TFBSs) from the UCSC TFBS; "CCBeder" is a set of attributes describing the GO characterisation of the query gene, also considering those of its

neighbours in the PPI. They were obtained by performing, for each gene set, a Fisher's exact test for the enrichment of interaction partners through the code made available by [22]. The GO-CC terms were downloaded from the g:Profiler web service as GMT files [23]. The log_{10} values of the p-values represent the features.

EGs and their functions are reported to be highly conserved. We added the number of orthologs for each query gene as a measure of conservation ("Orthologs count"), retrieved from [24].

In the context of cancer, EGs strongly overlap with cancer-driver genes, in which genetic mutations determining uncontrolled cellular growth occur with the highest frequency. For this reason, we collected renal cancer-specific data from DriverDBv3, which provides three types of drivers set according to the kind of alteration identified through multiple dedicated bioinformatics algorithms [25]. We counted, for each query gene, the times in which it is predicted as a Mutation driver ("Driver_genes_MUT"), a Copy Number Variation driver ("Driver_genes_CNV") or a Methylation driver ("Driver_genes_MET") for the three renal cancer (KIRCH, KIRP, KICH) or two lung cancer (LUAD, LUSC) subtypes, all together and separately. To characterise the genes according to their association with human diseases beyond cancer, we added the attribute "Gene-Disease association", in which the counts of these associations are reported as annotated in DisGeNet [26].

Data: Features' sets

The features described in the previous paragraph and those extracted through graph embedding were organised into four sets to be evaluated in the downstream experiments:

Bio: Kidney - 28 attributes, Lung - 25 attributes (consisting of all the features described in Table S1 except "CCcfs" and "CCBeder");

CCcfs: 3305 attributes (see Function and Localisation features in Table S1);

CCBeder: 1249 attributes (see Interaction features in Table S1);

N2V: 128 attributes, consisting of the node embeddings extracted as described in Section *Data: PPI embedding features*.

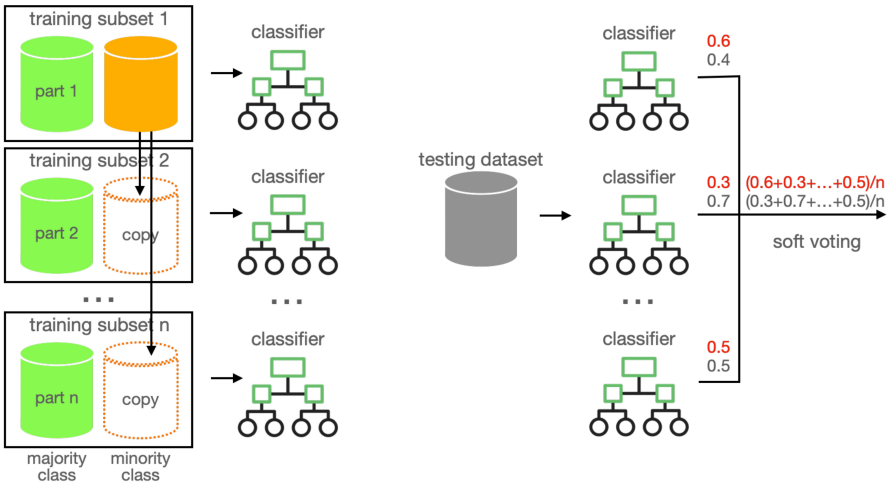
Feature data pre-processing

Gene feature data were pre-processed by removing attributes having constant values. In the case of the kidney context, we found two attributes with constant values in the Bio set ("Driver_genes_CNV (KICH)", "Driver_genes_MET (KICH)"), which are context-specific and related to tumour subtypes. In the case of the lung context, instead, none of the Bio attributes was constant. For both, three attributes with constant values were found in the CCcfs set ("Integral component of endoplasmic reticulum-Golgi intermediate compartment (ERGIC) membrane", "Intrinsic component of endoplasmic reticulum-Golgi intermediate compartment (ERGIC) membrane", "Pentameric IgM immunoglobulin complex").

Prediction model

The prediction model adopted in HELP has been specifically tailored to address the strong class imbalance issue inherently linked to EG prediction. We designed a new prediction model consisting of a soft-voting ensemble of n classifiers (see Fig 2). For training, we partition the set of majority class samples into n parts, and we train each classifier on a subset of training data composed of one of these parts along with the entire set of minority class training samples (Fig 2-(A)). For testing on unseen data,

each classifier of the ensemble produces a probability for the label prediction; we compute the final probability response of the ensemble as the average of the probabilities of the n voting classifiers (Fig 2-(B)). The number n of classifiers is chosen based on the class data distribution of training data so as to train each of the n classifiers on a balanced subset of data.



(A) (B)

Fig 2. Soft voting ensemble of n classifiers: (A) In the training stage the majority class samples are partitioned into n slices, each one coupled with the minority class samples to form datasets as inputs for the n trained models. (B) In the testing stage each new sample is predicted by each of the n classifiers and the final probability response is the average over the n partial responses.

In the experiments, the classifiers of the ensemble model are Light Gradient Boosting Machines (LightGBM), implemented in the scikit-learn Python library, and n has been chosen to ensure in the training subsets of data an imbalance ratio (IR), i.e., a ratio of the sample size of the majority class and that of the minority class, ranging from 1.2 and 1.6.

EG prediction results have been evaluated against the labels obtained using the two-class and three-class HELP labelling. The four binary prediction problems were validated via 5-fold stratified cross-validation. Each cross-validation round was iterated ten times, and we computed means and standard deviations of the collected metrics (described in Table S2) over the iterations.

Results

The purpose of the experimental study we conducted was threefold: 1) to ensure the effectiveness of the HELP labelling; 2) to find the optimal sets of gene attributes to construct a supervised predictor of gene essentiality based on the HELP labelling; 3) to analyse and discriminate the context-specific genes that show an intermediate behaviour between E and NE.

Comparison of HELP labelling with state-of-the-art cEGs

The HELP framework has been thought and created for identifying and predicting specifically the csEGs, even if it also allows the identification of the cEGs. To prove the

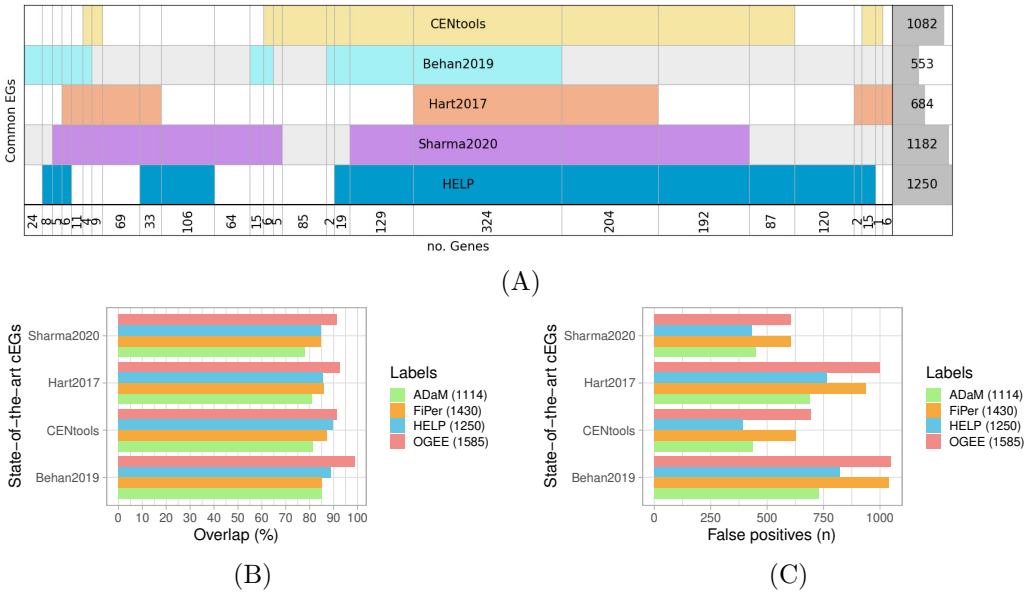


Fig 3. Comparison of HELP labelling with state-of-the-art cEGs. (A) Diagram representing the intersection of E genes labelled by HELP with state-of-the-art cEGs. Each row represents a different set of E genes. The last row reports the number of genes that result from the intersections. The last column on the right indicates the number of E genes for each set, with the dark grey shadow representing the corresponding histogram. (B) ADaM, FiPer, HELP and OGEE cEGs overlap with the state-of-the-art cEGs. In the legend, the number of cEGs identified by each approach is reported in brackets. (C) ADaM, FiPer, HELP and OGEE cEGs false positives, namely the genes not overlapping with the state-of-the-art cEGs. In the legend, the number of cEGs identified by each approach is reported in brackets.

effectiveness of the unsupervised HELP labelling, we intersected the cEGs computed as described in Section *Labelling* with some state-of-the-art sets used as reference in [11] (Fig 3(A)). As the diagram shows, the overlap among all the sets contains 324 genes. The HELP labelling agrees with CENtools [27] for 970 genes, Behan2019 [28] for 491, Hart2017 [29] for 584, and Sharma2020 [27] for 999 cEGs. The best overlap was therefore achieved in the case of the more recent sets of cEGs, i.e. CENtools and Sharma2020.

To further demonstrate the reliability of cEGs identification by HELP, we compared it with the approaches proposed by [11], implemented into the CoRe R package, and the list of cEGs provided by the OGEE database. The two methods of the CoRe package, the supervised Adaptive Daisy Model (ADaM) and the unsupervised Fitness Percentile (FiPer), were used on the CRISPR data matrix filtered on columns (cell lines) and rows (genes) as explained above. In the case of ADaM, as the method requires a binarised matrix, we first converted the input to binary scores by setting the threshold suggested in the package documentation ($\leq -0.5 = E$), then used the function "CoRe.PanCancer_ADaM". In the case of FiPer, we derived a consensus set of predicted E genes by intersecting the output of the three less stringent FiPer variants (Fixed, ROC-AUC, Slope). The comparison was performed through the evaluation of the overlaps with the state-of-the-art cEGs mentioned above (Fig 3-(B)). The OGEE list provided the greatest overlap with all four sets. OGEE considers common genes showing essentiality in at least 50% of cell lines. Such a low threshold determines a high

number of cEGS (1585), thus including many EGs of the reference sets but with a high number of false positives (Fig 3-(C)). The results by FiPer and ADaM confirmed what was commented by their authors regarding the stringency of the two methods (1114 ADaM, 1430 FiPer). The stringency grade influences the risk of erroneously identifying EGs. Indeed, FiPer detected much fewer cEGs than OGEE but with comparable false positives. ADaM instead, showed the best results in terms of false positives but with very low overlap percentages. In this scenario, HELP seems to provide the best compromise between the two evaluations. In two cases, it returned the lowest number of false positives (Sharma2020, CENtools) and was the second-best in terms of overlap percentages.

Context-specific EGs identification

We applied HELP to two tissue contexts, lung and kidney. The distributions of the obtained two-class and three-class labelling are reported in the left and right part of each column of Table 1, respectively.

Table 1. Distribution of the csEGs according to two-class and three-class HELP labelling. (A) kidney and (B) lung tissues. Left part of each column: distribution in the two classes E and NE. Right part of each column: distribution in the aE and sNE classes derived from NE. The count and percentage of labelled genes are reported.

(A) kidney

Class		Count		Percent	
E		1253		6.99%	
NE	aE	16678	3221	93.01%	17.96%
	sNE		13457		75.04%

(B) lung

Class		Count		Percent	
E		1235		6.89%	
NE	aE	16696	2849	93.11%	15.89%
	sNE		13847		77.22%

To evaluate the ability to identify the csEGs, we compared the obtained results with those identified by the CoRe's methods and the tissue-specific EGs obtained from the OGEE database. We used the "CoRe.CS_ADaM" function after binarising the input matrix and the FiPer consensus set (Fixed, ROC-AUC, Slope). Looking at the number of csEGs obtained (Fig 4), we can state that the most stringent sets are those provided by OGEE, even if, also in this case, the threshold on cell lines for considering a gene essential in a tissue is 50%. The higher stringency of ADaM compared to FiPer is not confirmed in the case of csEGs. The intersection sets between all the csEGs included 742 and 830 genes for kidney and lung, respectively. In both cases, the highest overlap is between HELP and ADaM (1192 and 1195 genes for kidney and lung, respectively). FiPer showed the highest number of genes not shared with the others (79) in the case of the kidney context, while for the lung, it was ADaM that provided the largest set of "unique" genes (200).

In our opinion, we cannot really talk about csEGs if we do not exclude from their computation the genes identified as cEGs, i.e. considering what we named ucsEGs. In the two cases considered in this work, using the HELP framework, we obtained 28 and 60 ucsEGs for Lung and Kidney contexts, respectively. The evident difference in numerosity could be ascribed partially to the different number of cell lines in the input

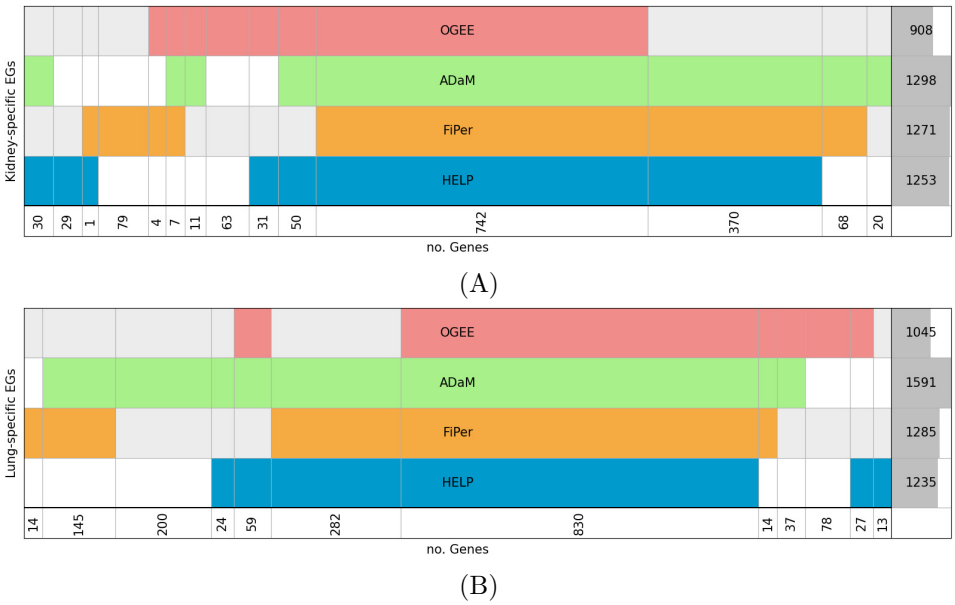


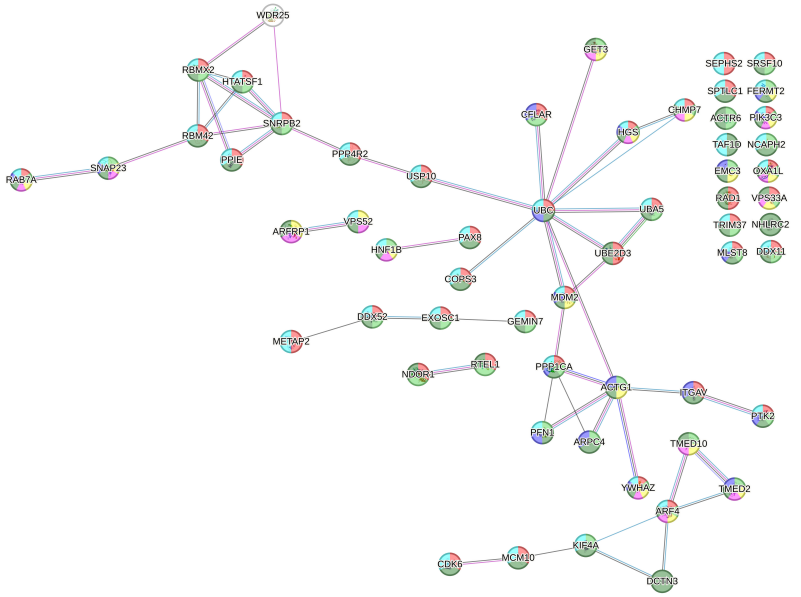
Fig 4. Comparison of HELP labelling with state-of-the-art csEGs. (A) kidney and (B) lung tissue. Diagram representing E genes intersections among ADaM, FiPer, OGEE and HELP labelling. Each row represents the set of E genes for a labelling. The last row reports the number of genes resulting from the intersections. The last column on the right indicates the number of E genes for each set, with the dark grey shadow representing the corresponding histogram.

data, making the results of the context with more data (i.e. lung) more stringent, but also to the different nature of the tissues, where the kidney can be somehow considered more heterogeneous. The enrichment analysis of the two sets (Fig 5) highlights that, besides the context-specificity, the EGs maintain their identity characteristics, such as the main localization in intracellular organelles, e.g. nucleus and mitochondrion, (GO Component "Intracellular membrane bounded organelle", "Intracellular organelle lumen"), or their involvement in growth functions (e.g. "Signal Transduction", "Cell cycle", "Cellular component biogenesis"). Most of them (41/60 in the kidney and 22/28 in the lung) contain the keyword "phosphoprotein" in their UniProt description, where protein phosphorylation is the key process to transmit extracellular and intracellular signals. Five genes were shared between the two tissues, suggesting that, besides the subtraction of the cEGs, some genes can be essential in different contexts and, then, rarely unique to a single one.

Optimal gene features towards essentiality prediction

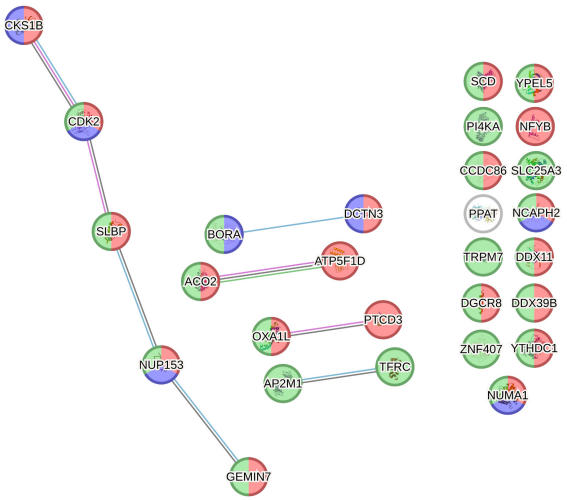
To proceed with the prediction task, we set for each classification problem the number n of classifiers of the soft voting ensemble model (Table 2) based on the classes sample size, as described in Section *Prediction model*. It should be observed that the number of labelled genes for each class (see Table 1) is generally slightly greater than the dimension of the corresponding set of genes adopted for prediction (see Table 2), as the latter excludes genes that do not have associated attributes adopted for prediction (11 genes).

We reported the prediction results (Fig 6 and S2) in terms of several metrics for evaluating the classification performance to allow easy comparisons, although aware that, especially in the case of unbalanced problems, the most reliable and informative



Category	Term ID	Term description	Count in network	FDR
GO Component	GO:0043231	Intracellular membrane-bounded organelle	55/12149	5.73E-05
GO Process	GO:0008152	Metabolic process	40/7988	0.0151
GO Process	GO:0071840	Cellular component organization or biogenesis	38/5639	9.22E-05
GO Process	GO:0008104	Protein localization	19/1943	0.0045
GO Process	GO:0015031	Protein transport	15/1180	0.0035
UniProt Keywords	KW-0597	Phosphoprotein	41/8122	0.0026
Reactome	HSA-162582	Signal Transduction	18/2540	0.0233

(A)



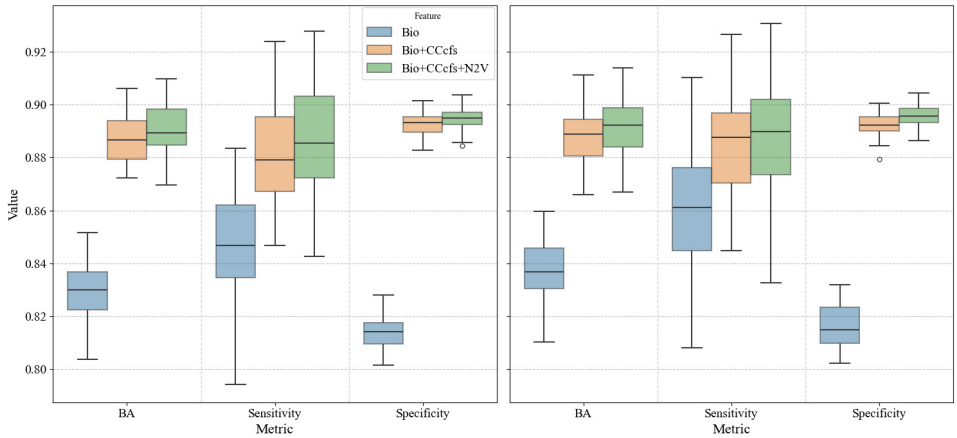
Category	Term ID	Term description	Count in network	FDR
GO Component	GO:0070013	Intracellular organelle lumen	20/5660	0.0075
UniProt Keywords	KW-0597	Phosphoprotein	22/8122	0.0219
Reactome	HSA-69278	Cell Cycle, Mitotic	7/526	0.0167

(B)

Fig 5. ucsEGs PPI enrichment. PPI networks built through STRING [30] using the ucsEGs computed for Kidney (A) and Lung (B). The nodes are coloured according to the enriched terms shown in the associated tables. The significant (False Discovery Rate, FDR < 0.05) non-redundant terms were ranked by the number of enriching genes (Count in the network: no. of enriching genes/no. of genes annotated for the term). The edges were built with all the STRING information except "Text mining".

Table 2. Number n of classifiers chosen for the soft voting ensemble prediction models for each binary classification problem (Problem) and tissue type (Kidney and Lung). IR (Imbalance Ratio) is the ratio of the sample size of the majority class (|majority|) and that of the minority class (|minority|) in the whole training dataset. IR_n indicates the IR in the training subset for each of the n classifiers.

Problem	IR (|majority|:|minority|)	n	IR_n
	Kidney		
E vs NE	12.87761 (15994:1242)	10	1.28824
E vs sNE	10.3752 (12886:1242)	8	1.29709
aE vs sNE	4.14607 (12886:3108)	3	1.38223
E vs aE	2.50241 (3108:1242)	2	1.2512
	Lung		
E vs NE	13.08006 (16010:1224)	10	1.308
E vs sNE	10.82597 (13251:1224)	8	1.35374
aE vs sNE	4.80283 (13251:2759)	3	1.60094
E vs aE	2.25407 (2759:1224)	2	1.12744



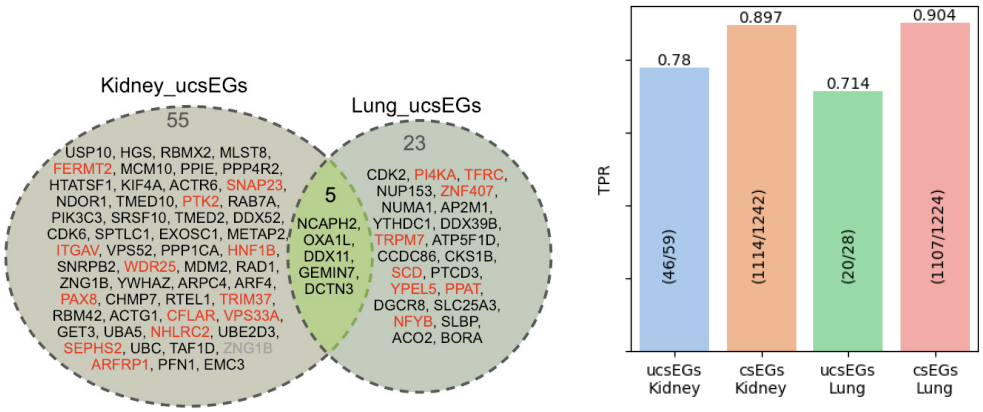
(A) (B)

Fig 6. Classification performance based on the two-class HELP labelling. Box plots of the performance metrics obtained for classification "E vs NE" using different sets of features on (A) kidney and (B) lung tissue.

one is the Balanced Accuracy (BA). Although the Bio set alone led, in all but the "aE vs sNE" problem, to the lowest performance results, it was included in each combination of features as it is the most varying set in terms of gene characteristics. The best-performing set among the multi-omics features was Bio+CCcfs+N2V for both contexts (Kidney - BA=0.890±0.009; Lung - BA=0.892±0.010). Unlike the others, the N2V set was analyzed alone since for some methods (e.g [31]) the PPI is used as the only input data for EGs classification. It was then added to the Bio+CCcfs combination to evaluate whether the network attributes could improve the performance. We can state that the embedding features produce a slight improvement with respect to the Bio+CCcfs set in both Sensitivity (Kidney: 0.886±0.019; Lung: 0.889±0.020) and Specificity (Kidney: 0.894±0.004; Lung: 0.895±0.005).

We concluded that, for both contexts, Bio+CCcfs+N2V was the optimal combination of gene features for the overall prediction of E and NE genes. In particular, the multi-omics attributes (Bio+CCcfs) strongly contributed to the capability of predicting the essentiality and the non-essentiality.

To evaluate the ability of the model to precisely predict the csEGs we calculated the True Positive Rates (TPRs) for ucsEGs and csEGs (Fig 7). ucsEGs of the two contexts and the intersection between the two, consisting of five genes, are shown in Fig. 7-(A), coloured according to the prediction results. The TPR for csEGs (Fig 7-(B)) corresponds to the Sensitivity, whose mean is reported in Fig S2, and was particularly high both for kidney and lung (0.897 and 0.904, respectively). The TPRs for ucsEGs (Fig 7-(B)) were instead lower due to the strong specificity of these small sets of genes. Nevertheless, achieving 78% (kidney) and 71% (lung) could be considered a significant outcome, suggesting promising predictive potential.



(A) (B)

Fig 7. ucsEGs prediction. (A) Diagram representing the intersection of ucsEGs of kidney and lung tissues. The ucsEGs for each tissue are computed by subtracting the cEGs calculated over all tissues from csEGs. ucsEGs wrongly predicted are highlighted in red. Only one kidney ucsEG (grey-coloured) was not predicted as no attributes were retrieved for it. (B) TPRs of prediction of ucsEGs and csEGs computed in one cross-validation round using the optimal features set Bio+CCcfs+N2V.

Although finding similar works in the literature to compare our results was difficult due to our focus on context-specificity, we considered some of them to generally evaluate the goodness of the measures and consolidate the effectiveness of HELP. In [22], the authors proposed an ML method across six model eukaryotes, called CLEARER, that on human cell lines obtained results comparable to HELP (ROC-AUC=0.955±0.018) using many more features (CLEARER: 41635; HELP: 3456 (Kidney) and 3455 (Lung) Bio+CCcfs+N2V). By embedding the human PPI network, [31] reported BA lower than ours (BA=0.783). DeepHE [4], which predicts EGs by integrating features derived from sequence data and PPI embedding, obtained a ROC-AUC=0.941. EPGAT [5], a method based on Graph Attention Networks, achieved a ROC-AUC=0.915±0.0034 using the combination of PPI and sublocalization annotations.

Investigating almost essentiality

One of the aims of the current work was to investigate the hypothesis of an intermediate class between E and NE, which we called aE. We analysed the classification performance on the three binary problems "E vs sNE", "E vs aE", and "aE vs sNE" (Fig 8; Figs S3-S5).

The combination Bio+CCcfs with and without the embedding features (N2V) led to the best results in terms of BA in all three problems, with the major contribution given by the biological part (Bio+CCcfs). As expected, discriminating E from sNE genes (Fig

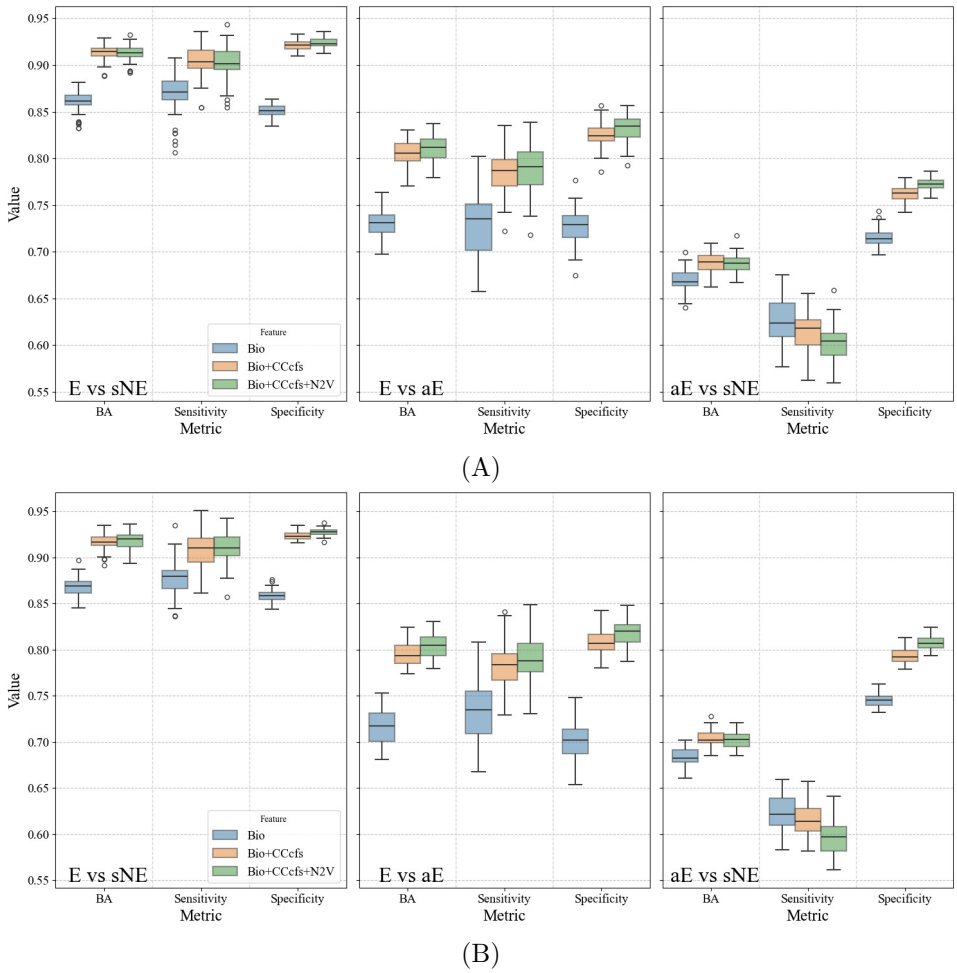


Fig 8. Classification performance based on the three-class HELP labelling. Box plots of the performance metrics obtained for classification of "E vs sNE", "E vs aE", and "aE vs sNE" using different sets of features on (A) kidney and (B) lung tissue.

8; Fig S3) resulted in better performance (BA=0.913±0.008 for kidney, BA=0.918±0.009 for lung) than considering NE all the genes that are not E (BA=0.890±0.009 for kidney, BA=0.892±0.010 for lung; see Fig S2). The most challenging problem was, as expected, the separation of the aE class from the other two (Fig 8; Figs S4-S5), given the intermediate nature of these genes. The separation of E and aE genes, although not optimal like E vs sNE, was achieved with a BA=0.811±0.013 for kidney, while BA=0.805±0.013 for lung. A more pronounced reduction of the performance was observed addressing the "aE vs sNE" problem (BA=0.688±0.011 for kidney, and BA=0.704±0.009 for lung), where the best BA is obtained with the Bio+CCcfs configuration of inputs. It is worth noticing how the best Sensitivity is obtained in the "aE vs sNE" problem when using only the bio attributes. However, as a counterpart, the Specificity is at a minimum.

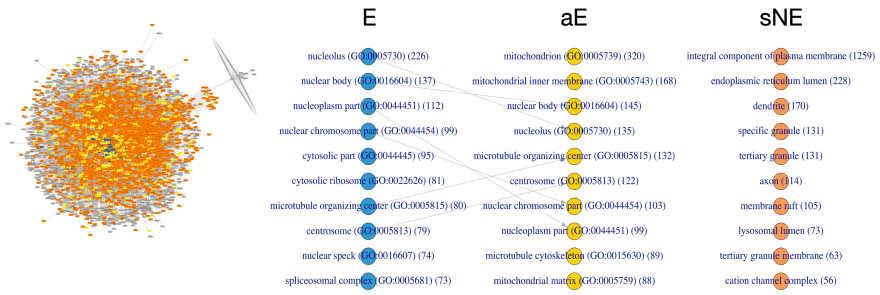
To get valuable insights about the three classes of genes, particularly the intermediate one aE, we investigated their profiles analysing the attributes that contribute to their discrimination (Bio+CCcfs+N2V). These genes showed a different localization in the PPI networks of both kidney and lung tissues (Fig 9, left side). As

the layout used to represent the networks is force-directed and tends to draw nodes with greater centrality to more central positions, the resulting figures strengthen the centrality-lethality rule. The E genes (blue nodes) essentially colocalize at the centre of the network, the aE genes (yellow nodes) mostly occupy the surrounding area, and the sNE genes (orange nodes) are distributed all around, with a sort of radial pattern. Looking at the GO-CC terms enrichment (Fig 9, right side), E genes mainly enriched the nuclear body, the aE genes the mitochondrion and, to a lesser extent, the nucleus, and the sNE genes the plasma membrane and intra- and extra-cellular transport systems (e.g. granule, lysosomes, membrane rafts). Furthermore, it was interesting to notice that while E and aE genes shared some terms, even with different ranking positions, none of them shared any term with the sNE genes. It is also worth noting that the Bio attributes were all but one statistically different between E and sNE genes, remarking their profound difference (Fig S1). Comparing the aE class with the other two, the situation did not change basically, confirming the hypothesis of three distinct groups, but some similarities were also present. E and aE genes, for example, showed comparable gene lengths and many Driver_Genes profiles. Similar trends were revealed by the context-specific attributes of the two tissues under study.

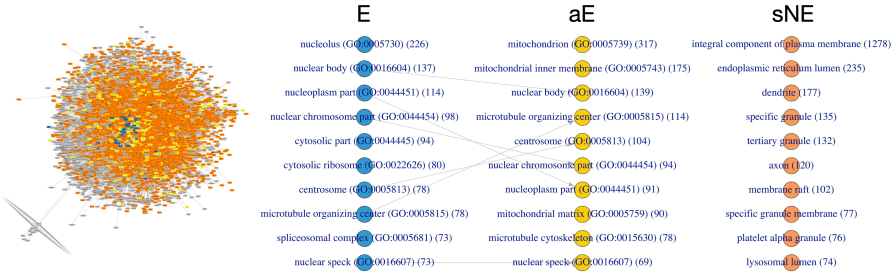
Discussion

This work was conceived to address two of the main issues concerning csEGs identification: labelling and prediction. The first regards the assignment of E/NE labels using the data furnished by gene deletion experiments. The only database we found providing tissue-specific lists of EGs is OGEE. Still, the need to create personalised contexts and the incessant updates of the experiments behind these lists require a flexible labelling method to apply to the knockout experimental scores. The second issue regards predicting those labels by training ML methods with gene attributes from multi-source data. The labelling task can be considered an identification method, but it strictly depends on generating experimental scores. Instead, the ML approaches are conceived to support or even substitute costly and timely intensive wet lab procedures.

We presented the methods and results focusing on two human tissue contexts to demonstrate the effectiveness of the HELP framework. We applied an unsupervised thresholding method to derive both cEGS and csEGs from CRISPR scores. The effectiveness of the obtained labels was assessed by comparing the resulting cEGS with the CoRe's methods, ADaM and FiPer, and with four state-of-the-art cEGs. The overlap with these sets was particularly high, especially in the case of the most recent ones. The results make HELP the best compromise between the overlap with reference sets and false positive rates, with the advantage over FiPer being more stringent and over ADaM being unsupervised. Although HELP includes the identification of cEGs, this step was mainly designed and implemented to capture the context specificity by subtracting the genes that are essential in most contexts. This subtraction step gives rise to what we call ucsEGs. The stringency is crucial to guarantee higher confidence when approaching context-specific EGs. The discrimination of the "essentiality" classes was performed by training a prediction model through knowledge from multi-omics data and embedding a context-specific PPI. As evident, the classification faces a strong unbalance issue, where the majority class can be 13 times bigger than the minority one. This strongly affects the classification performance, returning high Specificity and low Sensitivity. To this extent, we adopted a soft-voting strategy on an ensemble of classifiers, whose number was given by partitioning the majority class set based on the data distribution. This approach reduced the unbalancing, still using the whole dataset, in contrast to the subsampling, without duplicating data, as done by the oversampling, and without introducing synthetic data, as in the augmentation. Different combinations



(A)



(B)

Fig 9. Cellular localization of E, aE and sNE genes. The cellular sublocation of the three classes of genes is shown through the PPI network representation (left side) and the top-ranked enriched GO-CC terms (right side) for kidney (A) and lung (B) contexts. E, aE and sNE genes are highlighted in blue, yellow and orange, respectively, in both graphics. In favour of visualisation, only the largest connected component of the network is shown. Grey nodes are genes in the PPI not labelled by HELP since not included in the experimental data. The network was plotted through Cytoscape v3.10.1 [32]. The rank-plots (right-side) show the top 10 significant terms ordered by gene count. The number of genes enriching each term is in brackets. The connection lines indicate the terms shared by the groups. The enrichment analysis was performed using the enrichR package v3.2 [33].

of the gene attributes were used until the best-performing configuration was given by combining two sets of features, Bio+CCcfs, in some cases with the addition of the embedding features N2V. The Bio set contains a collection of generic and tissue-specific functional and structural characteristics; the CCcfs set contains information on cellular localisation and involvement in molecular complexes; the N2V set contains topological information from the PPI. The model could correctly predict around 90% of csEGs and between 70%-80% of ucsEGs. This discrepancy was likely due to the inherent challenge in modelling the context-specificity of EGs, stemming from their small number compared to the csEGs and the limited availability of tissue-specific information for characterizing them accurately.

Several recent papers discussing the characteristics of EGs [9, 34] reveal that essentiality is not a binary fixed but a flexible status depending on the genetic and environmental contexts. Furthermore, besides genes having a direct role in growth and viability, a set of genes is involved in essential functions and interactions [35]. Ultimately, the characteristics predictive of the essentiality are quantitative traits, as are the essentiality scores. In this scenario, it is reasonable to think that the dividing

line between essentiality and not is uncertain and that a gene essential in some contexts likely keeps involvement in essential functions in others. In our previous works, using a different labelling strategy based on a knowledge-driven subdivision of CRISPR scores (CS), we identified the best configuration of E and NE genes by ML trials. We achieved the best performance training a model on biological and embedding attributes (CS0 (E) vs CS6-9 (NE): BA=0.844 [36]; CS0 (E) vs CS6-9 (NE): BA=0.846 [37]). The investigation of the intermediate groups led us to think about possibly overcoming the dichotomic view of essentiality. The experimental evidence we got, along with the above considerations, suggest the existence of shades of "essentiality", that, in the most simplistic view, can be represented by a third class of "almost Essentials" (aE). We consistently improved performance by excluding the aE genes from the NE class. The prediction of the aE genes was the most challenging task, especially in the case of the "aE vs sNE" problem. Nonetheless, looking at the biological characteristics, the existence of a group of genes different from NE and E but more similar to the latter was evident. This result was particularly interesting given that the aE class was extracted from the NE class, remarking the weakness of the binary separation. As the Bio+CCcfs, eventually +N2V, combination determined the best performance, we analysed these features to learn about the differences among the three classes. The features in the Bio set were statistically different in all but one of the cases comparing E and sNE, and in 29 cases over 32, summing the comparisons aE vs E and aE vs sNE of both lung and kidney contexts. Particularly interesting was the enrichment of the GO-CC terms for the three sets of genes. Ranking the significant terms by gene count, it was worth noticing that no terms were shared between the two essential groups, E and aE, and sNE genes. Furthermore, the enriched terms gave an exact idea of the differences and similarities between these groups. E genes mostly enriched the nucleus area, aE genes the mitochondrion first and the nucleus. This result highlights the less pronounced border between E and aE, and the particularly interesting role of the aE as candidate biomarkers, as the mitochondria are considered not only powerhouses but also dynamic regulators of life, death, proliferation, motion and stemness of cancer cells [38]. Moreover, this perfectly aligns with the role of mitochondria to fill the functional gap between the nucleus and cytoplasmic organelles and the interconnection between their components [39]. In contrast, sNE genes enriched the plasma membrane and cellular transportation systems. This different localisation was also evident by visualizing the three groups in the PPI networks, which seems to reproduce the picture of a human cell, with the E genes localised in the core of the network, aE in the surrounding area and sNE widely spread in the cytoplasm and at the borders. Our recent work demonstrated that even integrating the PPI with a metabolic network to add a functional centrality to the physical one, the contribution was totally in charge of the PPI [37]. This can likely be explained by the fact that while the metabolic machinery comprises several alternative paths to achieve a specific objective, the lack of a component involved in many physical complexes and interactions is hard to tolerate.

We believe that these promising insights open the way to future work to investigate the essentiality shades further and apply HELP to examine cancer-specific EGs.

Supporting information

Data/Software Availability All data and software are released under the GNU licence and are available at: https://doi.org/10.5281/zenodo.10964744. Here you can find the HELPpy package implementing the HELP framework in Python, with a few jupyter notebooks exemplifying the identification of csEGs (`labelling.ipynb`) and ucsEGs (`csegs.ipynb`), the prediction of csEGs (`prediction.ipynb`), the extraction of PPI embeddings (`embedding.ipynb`) and the reproduction of the prediction

experiments (`experiment.ipynb`).

Fig S1. Boxplots of the Bio attributes values for the E, aE, and sNE classes. The stars on the top indicate the significance of the Wilcoxon test for each pair of comparisons (**** ≤ 0.0001, *** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05, ns = not significant). The other Driver genes attributes were not shown as having small ranges of values and poor statistics.

Fig S2. "E vs NE" classification performance based on HELP labelling. (A) kidney tissue, (B) lung tissue.

Fig S3. "E vs sNE" classification performance based on HELP labelling. (A) kidney tissue, (B) lung tissue.

Fig S4. "E vs aE" classification performance based on HELP labelling. (A) kidney tissue, (B) lung tissue.

Fig S5. "aE vs sNE" classification performance based on HELP labelling. (A) kidney tissue, (B) lung tissue.

Table S1. Collected genomic, transcriptomic, epigenetic, functional and evolutionary features of genes. (cs) indicates the context-specific attributes.

Table S2. Classification performance metrics adopted in the experiments. They are defined in terms of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), where the first class in each task (e.g. class E in the "E vs NE" classification task) is assumed as the positive class.

Acknowledgments

This work was carried out also within the activities of the authors as members of the INdAM Research group GNCS and the ICAR-CNR INdAM Research Unit. Authors from ICAR-CNR would like to thank R. Mattiello, G. Trerotola and S. Sada for their technical support.

References

1. Ferreira P, Choupina AB. CRISPR/Cas9 a simple, inexpensive and effective technique for gene editing. Mol Biol Rep. 2022;49(7):7079–7086.
2. Aromolaran O, Aromolaran D, Isewon I, et al. Machine learning approach to gene essentiality prediction: a review. Brief Bioinform. 2021;22(5):bbab128.
3. Giordano M, Falbo E, Maddalena L, Piccirillo M, Granata I. Untangling the Context-Specificity of Essential Genes by Means of Machine Learning: A Constructive Experience. Biomolecules. 2024;14(1).
4. Zhang X, Xiao W, Xiao W. DeepHE: Accurately predicting human essential genes based on deep learning. PLoS Comput Biol. 2020;16(9):e1008229.
5. Schapke J, et al. EPGAT: Gene Essentiality Prediction With Graph Attention Networks. IEEE/ACM Trans Comput Biol Bioinform. 2022;19(3):1615–1626.

6. Le NQK, Do DT, et al. A computational framework based on ensemble deep neural networks for essential genes identification. Int J Mol Sci. 2020;21(23):9070.

7. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, et al. A systematic survey of centrality measures for protein-protein interaction networks. BMC Syst Biol. 2018;12(1):1–17.

8. Larrimore KE, Rancati G. The conditional nature of gene essentiality. Curr Opin Genet Dev. 2019;58:55–61.

9. Rancati G, Moffat J, Typas A, Pavelka N. Emerging and evolving concepts in gene essentiality. Nat Rev Genet. 2018;19(1):34–49.

10. Gurumayum S, et al. OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. Nucleic Acids Res. 2021;49(D1):D998–D1003.

11. Vinceti A, Karakoc E, et al. CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. BMC Genom. 2021;22(828).

12. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans Syst Man Cybern. 1979;9(1):62–66.

13. Kotlyar M, Pastrello C, Ahmed Z, et al. IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis. Nucleic Acids Res. 2021;50(D1):D640–D647.

14. Grover A, Leskovec J. Node2vec: Scalable Feature Learning for Networks. In: KDD '16; 2016. p. 855–864.

15. Mikolov T, Sutskever I, et al. Distributed Representations of Words and Phrases and Their Compositionality. In: Proc. NIPS'13 - Vol.2; 2013. p. 3111–3119.

16. Ardlie KG, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):648–660.

17. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216–W221.

18. Tang G, Cho M, Wang X. OncoDB: an interactive online database for analysis of gene expression and viral infection in cancer. Nucleic Acids Res. 2022;50(D1):D1334–D1339.

19. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. Science. 2015;347(6220).

20. Binder JX, Pletscher-Frankild S, Tsafou K, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database. 2014;2014.

21. Oughtred R, Stark C, Breitkreutz BJ, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):D529–D541.

22. Beder T, Aromolaran O, Dönitz J, et al. Identifying essential genes across eukaryotes by machine learning. NAR Genom Bioinform. 2021;3(4):lqab110.

23. Reimand J, Kull M, et al. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007;35(suppl_2):W193–W200.

24. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res. 2023;51(D1):D29–D38.

25. Liu SH, Shen PC, Chen CY, et al. DriverDBv3: a multi-omics database for cancer driver gene research. Nucleic Acids Res. 2020;48(D1):D863–D870.

26. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845–D855.

27. Sharma S, Dincer C, Weidemüller P, et al. CEN-tools: an integrative platform to identify the contexts of essential genes. Mol Syst Biol. 2020;16(10):e9698.

28. Behan FM, Iorio F, Picco G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568(7753):511–516.

29. Hart T, Tong AHY, Chan K, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 Genes—Genomes—Genetics. 2017;7(8):2719–2727.

30. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research. 2019;47(D1):D607–D613.

31. Dai W, Chang Q, Peng W, Zhong J, Li Y. Network embedding the protein–protein interaction network for human essential genes identification. Genes. 2020;11(2):153.

32. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504.

33. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–W97.

34. Chen H, Zhang Z, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Brief Bioinform. 2020;21(4):1397–1410.

35. Lachance JC, Matteau D, Brodeur J, et al. Genome-scale metabolic modeling reveals key features of a minimal gene set. Mol Syst Biol. 2021;17(7):e10099.

36. Manzo M, Giordano M, et al. Novel Data Science Methodologies for Essential Genes Identification Based on Network Analysis. In: Data Science in Applications. Springer; 2023. p. 117–145.

37. Granata I, Giordano M, et al. Network-Based Computational Modeling to Unravel Gene Essentiality. In: Mondaini RP, editor. Trends in Biomathematics: Modeling Epidemiological, Neuronal, and Social Dynamics. Springer Nature Switzerland; 2023. p. 29–56.

38. Grasso D, Zampieri LX, Capelôa T, et al. Mitochondria in cancer. Cell stress. 2020;4(6):114.

39. Gasparotto M, et al. Nuclear and Cytoplasmatic Players in Mitochondria-Related CNS Disorders: Chromatin Modifications and Subcellular Trafficking. Biomolecules. 2022;12(5):625.