nature computational science

Article

https://doi.org/10.1038/s43588-024-00733-1

Comprehensive prediction and analysis of human protein essentiality based on a pretrained large language model

Received: 10 April 2024

Accepted: 31 October 2024

Published online: 27 November 2024

Check for updates

Boming Kang [1,3], Rui Fan [1,3], Chunmei Cui[1] & Qinghua Cui [1,2]

Human essential proteins (HEPs) are indispensable for individual viability and development. However, experimental methods to identify HEPs are often costly, time consuming and labor intensive. In addition, existing computational methods predict HEPs only at the cell line level, but HEPs vary across living human, cell line and animal models. Here we develop a sequence-based deep learning model, Protein Importance Calculator (PIC), by fine-tuning a pretrained protein language model. PIC not only substantially outperforms existing methods for predicting HEPs but also provides comprehensive prediction results across three levels: human, cell line and mouse. Furthermore, we define the protein essential score, derived from PIC, to quantify human protein essentiality and validate its effectiveness by a series of biological analyses. We also demonstrate the biomedical value of the protein essential score by identifying potential prognostic biomarkers for breast cancer and quantifying the essentiality of 617,462 human microproteins.

Essential genes are key components of the minimal genome required for an organism's survival, making them indispensable[1]. Essential proteins encoded by these genes often perform core biological functions related to growth and development[2]. Thus, understanding protein essentiality is crucial for identifying drug targets, clinical therapeutics and applications in synthetic biology. Experimental methods such as single gene deletion, RNA interference and clustered regularly interspaced short palindromic repeats (CRISPR) gene-editing technologies are commonly used to identify essential proteins[3]. However, these approaches are often costly, time consuming, labor intensive and dependent on specific types. Therefore, there is an urgent need to develop precise and efficient computational methods for assessing protein essentiality.

Up to now, several computational approaches have been developed and can be broadly classified into two main categories: network-based methods and sequence-based methods[3]. Network-based methods often rely on a protein–protein interaction (PPI) network and use metrics such as betweenness centrality[4], closeness centrality[5] and degree centrality[6] to assess the importance of each protein[7]. However, these methods are highly dependent on the PPI networks, which often have serious bias because they are unable to analyze proteins that lack PPI information[3]. In contrast, sequence-based methods are more accessible due to the central role of the sequence–structure–function paradigm. These sequence-based methods commonly utilize integrated sequence intrinsic features or biometric information to represent nucleic acid or protein sequences and then predict essentiality via machine learning models. Recently, a series of sequence-based models have been proposed, including Pheg[8], DeeplyEssential[9], DeepHE[10], EP-GBDT[11], EP-EDL[12] and DeepCellEss[13]. However, the performances of these methods are limited by the quality of features extracted from sequences, due to the complexity and heterogeneity of protein sequences.

More recently, large language models have achieved notable success in the field of natural language processing[14]. Inspired by large language models, protein language models (PLMs) pretrained on large-scale protein sequences have emerged, including Evolutionary

[1]Department of Biomedical Informatics, State Key Laboratory of Vascular Homeostasis and Remodeling, School of Basic Medical Sciences, Peking University, Beijing, China. [2]School of Sports Medicine, Wuhan Institute of Physical Education, Wuhan, China. [3]These authors contributed equally: Boming Kang, Rui Fan. e-mail: cuiqinghua@bjmu.edu.cn

https://doi.org/10.1038/s43588-024-00733-1

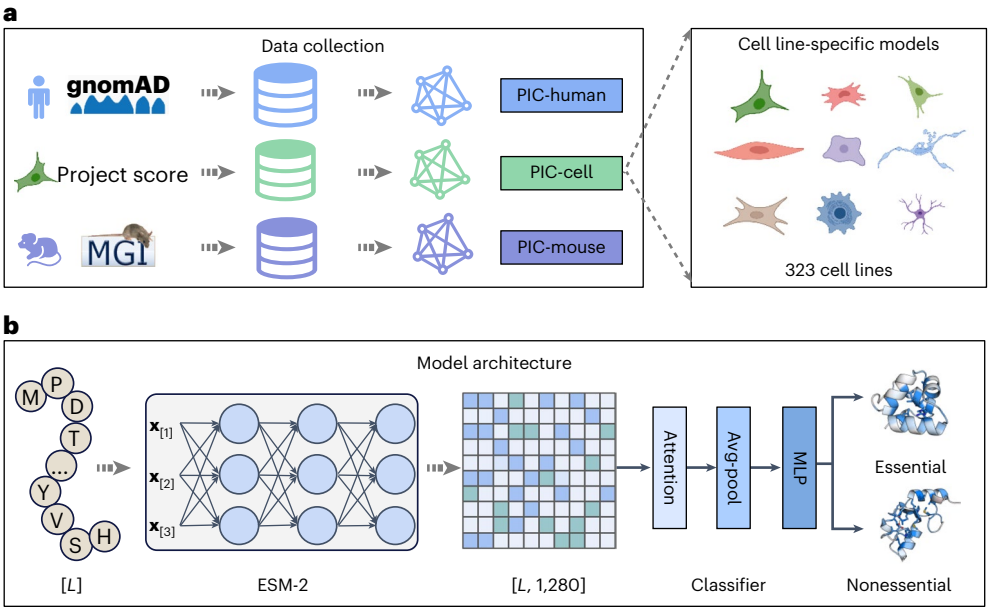


Fig. 1 | Overall workflow. a, Data collection. Human protein essentiality data were collected from gnomAD, Project Score and MGI databases to construct PIC-human, PIC-cell and PIC-mouse models, respectively. PIC-cell model is a cell line-specific model, including a total of 323 different human cell lines. **b**, Model architecture. All PIC models had an identical architecture, including three main modules: Embedding, Attention and Prediction. The Embedding module extracts input protein sequence features using ESM-2, the Attention module captures the relative importance of amino acids at different positions within the protein sequences, and the Prediction module determines the predicted essential probability of the input protein sequence. Avg-pool is a method used for calculating the average token embeddings in a sequence to generate a unified representation of the entire sequence. MLP, multi-layer perceptron. The subfigure on the right panel of **a** was created in BioRender. Kang, B. (2024) https://BioRender.com/c12h120.

Scale Modeling (ESM)[14] and ProtTrans[15]. PLMs provide a better representation of protein sequences by embedding the protein sequences in the form of high-dimensional numerical vectors. With the protein representations captured by the PLMs, the performances on diverse downstream tasks, such as structure prediction[14], subcellular localization prediction[16], signal peptide prediction[17] and N-linked glycosylation sites prediction[18], have been revolutionized. However, it remains unknown whether PLMs can substantially improve the task of protein essentiality prediction.

Current protein essentiality prediction models commonly collect data from cell viability assays for training, where essential proteins are defined as a substantial decrease in cell viability after their knockout. However, human protein essentiality is context dependent and closely related to cell type and cellular physiological stage. Yet, current methods, except DeepCellEss, commonly overlooked these differences and only developed general models based on integrated essentiality data rather than cell type-specific data. Besides, human essential proteins can also be identified by human genome sequencing projects and animal embryonic lethality assays like in mice. At the human level, proteins are defined as essential if they are rarely disturbed or truncated because these proteins are more intolerant to loss-of-function variants. Meanwhile, at the mouse level, essential proteins are defined as human–mouse homologs that cause embryonic death in mice after being knocked out. In addition, previous studies have demonstrated that proteins identified as essential in human cell lines or knockout mice may be distinct from those in living humans[1], highlighting the urgent need for a comprehensive prediction and analysis of the human essential proteins.

Hence, for comprehensively and systematically assessing human protein essentiality across three levels (human, mouse and cell line), we proposed a deep learning-based method, Protein Importance Calculator (PIC), by fine-tuning PLMs, which achieved state-of-the-art performance on the human protein essentiality prediction task compared with existing methods. Moreover, based on the probability values output by PIC models, we defined the protein essential score (PES) as a metric to quantify the protein essentiality and performed a series of cross-level analyses to validate its biological meaning. Furthermore, we confirmed the power of PES by utilizing it to identify potential prognostic biomarkers of breast cancer and quantify the essentiality of human microproteins. Finally, we developed a user-friendly web server (http://www.cuilab.cn/pic) for the convenience of researchers.

Results

Overview of PIC models

PIC is a series of deep learning models designed for comprehensive prediction of human essential proteins, including a total of 325 PIC models across three different levels: one model for the human level (PIC-human), one for the mouse level (PIC-mouse) and 323 models for the cell line level (PIC-cell). Protein essentiality data were collected from gnomAD[19], OGEE-MGI[20] and Project Score databases[21] to train PIC-human, PIC-mouse and PIC-cell, respectively (Fig. 1a). All PIC models shared an identical architecture comprising three main modules: Embedding, Attention and Prediction (Fig. 1b). For the 323 cell-level PIC models, we used a soft voting strategy within an ensemble learning framework[22] to aggregate 323 cell-level PIC models' prediction results, resulting in the high-performing PIC-cell model. In addition, ensemble learning was applied to develop 28 disease-level PIC models (Supplementary Table 1) and 19 tissue-level PIC models (Supplementary Table 2), enabling prediction of human protein essentiality in specific diseases or tissues. To optimize the PIC model architecture, we conducted a series of ablation studies and hyperparameter optimization experiments. The results led us to select the ESM-2 model with 650 million parameters for protein sequence feature extraction, applying an average pooling method to generate representations for complete protein sequences. (Fig. 2, Supplementary Note 1and Supplementary Data 1).

Overall performance of PIC models

We evaluated the performance of PIC models on their respective independent test datasets using metrics including Accuracy, Recall, Precision, F1 score, area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC)

https://doi.org/10.1038/s43588-024-00733-1

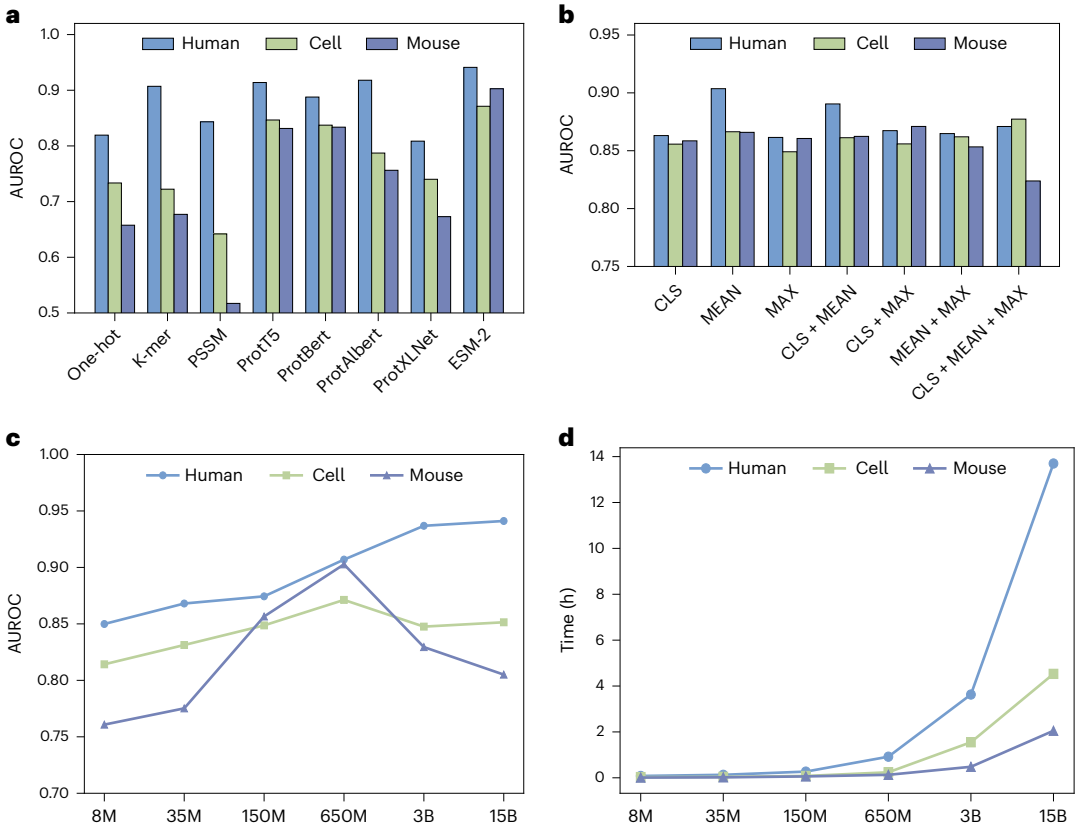


Fig. 2 | Ablation studies of PIC models. a, The AUROC values for different feature extraction methods are compared across PIC-human, PIC-cell and PIC-mouse models. Methods include One-hot, K-mer, PSSM, ProtT5, ProtBert, ProtAlbert, ProtXLNet and ESM-2. **b**, The AUROC values for different pooling strategies are compared across PIC-human, PIC-cell and PIC-mouse models. Strategies include CLS, MEAN, MAX and their combinations. **c**, The AUROC values for PIC-human, PIC-cell and PIC-mouse models with varying sizes of the model parameters. **d**, The training time (in hours) for PIC-human, PIC-cell and PIC-mouse models with varying sizes of the model parameters. The data for PIC-cell models are presented as mean values in the figure. M, million; B, billion. CLS: using the special [CLS] token's embedding from the output of ESM-2 to represent the entire protein sequence. MEAN: using the average of all token embeddings to represent the entire protein sequence. Max: selecting the maximum value from each dimension of the token embeddings to represent the entire protein sequence.

(Supplementary Tables 3 and 4). PIC-human achieved the highest AUROC at 0.9132, followed by PIC-mouse with an AUROC of 0.8736. The KYSE-70 cell-level model, whose AUROC (0.8579) is the median of the 323 cell-level PIC models, was selected to represent the average performance of PIC-cell models (Fig. 3a). To further assess PIC model performance, we compared it against three widely used open-source sequence-based protein essentiality prediction models. Among the models compared, EP-EDL and EP-GBDT were trained on integrated datasets from cell viability assays, while DeepCellEss is a cell line-specific model based on data from 323 human cell line datasets. Moreover, we designed PIC-base as a self-baseline model, which uses sequence-level feature vectors directly output by ESM-2 for protein essentiality prediction. The results show that PIC increases AUROC by 5.13–12.10% and also substantially improves Accuracy, Precision, F1 score and AUPRC (Supplementary Table 5) compared with existing methods. Given that DeepCellEss is cell line specific, we further compared AUROC and AUPRC values between PIC and DeepCellEss across 323 cell lines individually. As a result, compared with DeepCellEss, PIC improves the AUROC and AUPRC by an average of 9.64% and 10.52% on 323 cell lines, respectively (Fig. 3b,c and Supplementary Table 6). Moreover, the AUROC values of the 19 tissue-level and 28 disease-level PIC-cell models ranged from 0.7543 to 0.9029 (Fig. 3d,e and Supplementary Tables 1 and 2).

Biological relevance of PES generated by PIC

As presented in Methods, we defined PES output by PIC models to measure human protein essentiality. A previous study[23] reported that essential proteins are commonly correlated with other biological metrics characterizing protein importance. Therefore, we performed Spearman correlation analysis between PES and these well-established biological metrics, including PPI network node degree, normal tissue expression level, cancer tissue expression level, PhyloP, PhastCons and the number of related diseases. PES has strong positive correlations with the protein degree in the PPI network ($\rho = 0.4046$, $P < 1.0 \times 10^{-323}$), normal tissue expression level ($\rho = 0.3095$, $P < 1.0 \times 10^{-323}$), cancer tissue expression level ($\rho = 0.4745$, $P < 1.0 \times 10^{-323}$), PhyloP ($\rho = 0.4585$, $P < 1.0 \times 10^{-323}$), PhastCons ($\rho = 0.4081$, $P < 1.0 \times 10^{-323}$) and the number of related diseases ($\rho = 0.1912$, $P = 9.3 \times 10^{-139}$) (Fig. 4a). In addition, we ranked human proteins by PES in ascending order and then divided them into ten equal-sized groups to observe the trend of various biological metrics from group 1 to group 10. The results also showed robust positive correlations between PES and all of the observed biological metrics (Fig. 4b and Supplementary Figs. 1–6).

It is well known that essential proteins tend to perform crucial biological functions at corresponding subcellular localization. We further grouped human proteins according to their main subcellular localization categories based on the Human Protein Atlas database[24] annotation and then calculated PES for proteins in each category. We found that proteins with higher essentiality tend to localize closer to the cell nucleus, while those closer to the cell membrane tend to have lower PES (Fig. 4c). Furthermore, we divided proteins into two groups based on whether they are localized at the cell nucleus. Then, we conducted a chi-square test for all human proteins between the binary essentiality

https://doi.org/10.1038/s43588-024-00733-1

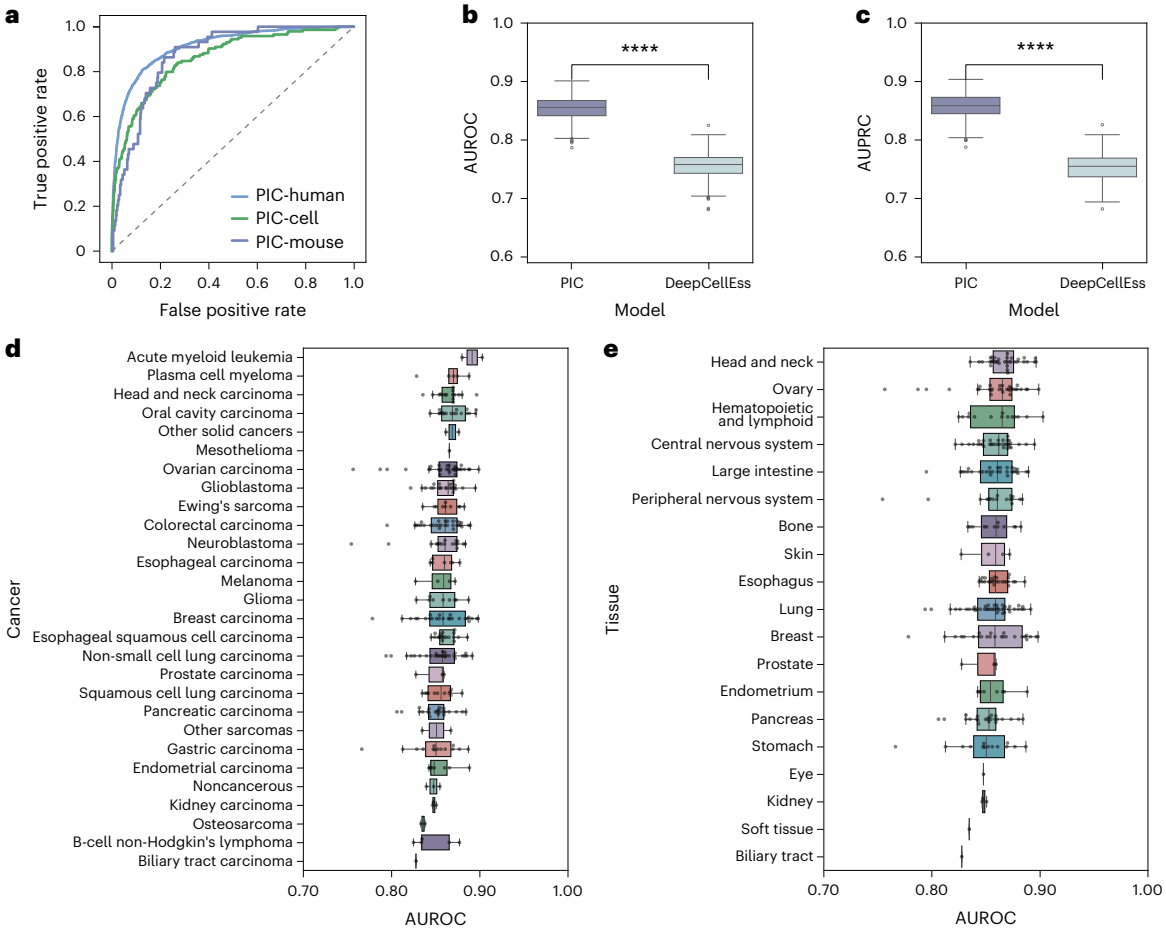


Fig. 3 | Performance presentation and comparison of PIC models. a, ROC curves of PIC-human ($n = 6{,}505$), PIC-mouse ($n = 1{,}718$) and PIC-cell models ($n = 627$), where n represents the number of samples used to generate each ROC curve. The data for PIC-cell models are presented as mean values in the figure. **b**, Comparison of AUROC between PIC and DeepCellEss across 323 human cell lines ($P = 3.18 \times 10^{-231}$). **c**, Comparison of AUPRC between PIC and DeepCellEss on 323 human cell lines ($P = 5.36 \times 10^{-225}$). **d**, AUROC values of PIC-cell models for 28 cancer types. **e**, AUROC values of PIC-cell models for 19 tissue types. ****$P < 0.0001$. The box plots show the median (center line), the 25th and 75th percentiles (bounds of the box) and the minimum and maximum values (whiskers). P values were calculated using a paired-sample t-test (two-sided, $n = 323$, where n represents the number of cell lines).

defined by PES and the binary subcellular localization. The results showed that essential proteins tend to locate at the cell nucleus, while nonessential proteins tend to locate outside the cell nucleus (chi-square test, $P = 3.42 \times 10^{-66}$). These observations are consistent with the fact that vital cellular processes such as replication, transcription and translation tend to occur near the cell nucleus.

We next investigated the relation between PES and protein functions. We first sorted human proteins by PES in ascending order and then divided them into ten equal-sized bins. Subsequently, we performed functional enrichment analyses on proteins within each bin, including Biological Process (BP), Cellular Component (CC), Molecular Function (MF) and Kyoto Encyclopedia of Genes and Genomes (KEGG). Then, we aggregated annotation terms from the ten bins, focusing on terms shared by all bins to observe relations between functions and essentiality. The odds ratio (OR) was calculated for each annotation term and then transformed via $\log_2$, denoted as $\log_2$OR. Spearman correlation coefficients were then calculated between each term's $\log_2$OR and the bin number across ten bins. We selected the top terms with the highest positive and the negative correlation coefficients to show the relation between functions and essentiality. Proteins with higher PES tend to be more associated with essential biological processes such as cell division, DNA biosynthetic process and peptide biosynthetic process. In contrast, proteins involved in more complex functions such as humoral immune response, neuropeptide signaling pathway and fatty acid metabolic process tend to have lower PES (Fig. 4d and Supplementary Figs. 7 and 8). As for KEGG, we observed that highly essential proteins tend to be linked to neurodegenerative diseases such as Alzheimer's disease, Huntington's disease and prion disease, while highly nonessential proteins tend to be connected to functions related to digestion and metabolism, including protein digestion and absorption, drug metabolism and cholesterol metabolism (Fig. 4e and Supplementary Data 2 and 3). As the PES increases, protein functions are more associated with essential biological processes such as DNA replication, protein synthesis and cell division. Conversely, as the PES decreases, protein functions tend to be more involved in complex biological roles such as immune responses, nutrient metabolism and nervous system maintenance.

Cross-level analyses based on PES at different levels

Protein essentiality varies across different levels (human, cell line and mouse) and also shows variability among different cell lines. To explore this, we first calculated the number of essential genes identified by wet-lab experiments in each cell line and found substantial variation in the number of essential genes across the 323 cell lines, ranging from 353 to 2,117 (Fig. 5a). Moreover, due to the variation of protein essentiality across living humans, cell lines and mice, we defined hPES, cPES and mPES to measure protein essentiality at the human, cell line and mouse levels, respectively. We found that mPES and cPES exhibits a

https://doi.org/10.1038/s43588-024-00733-1

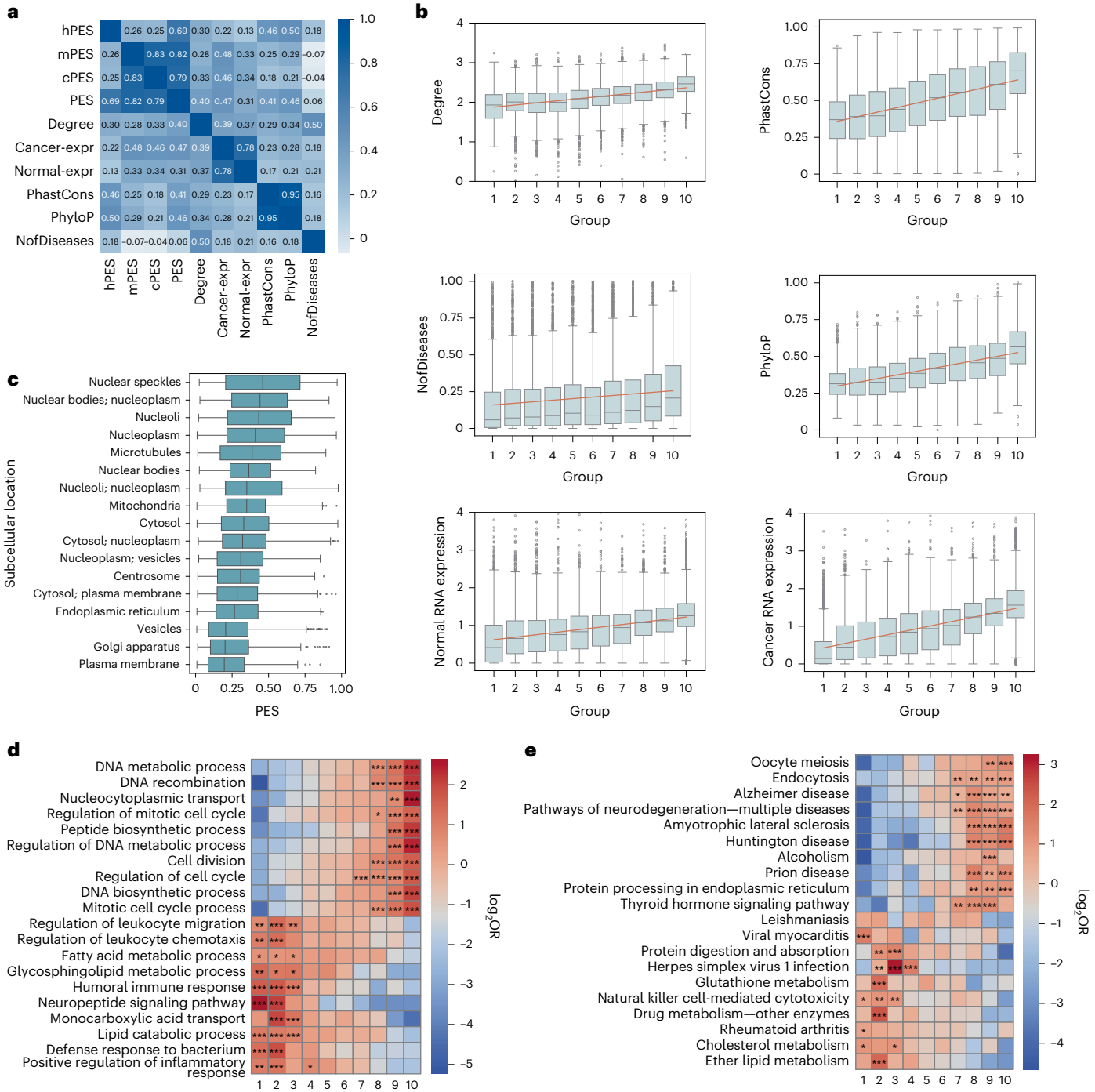
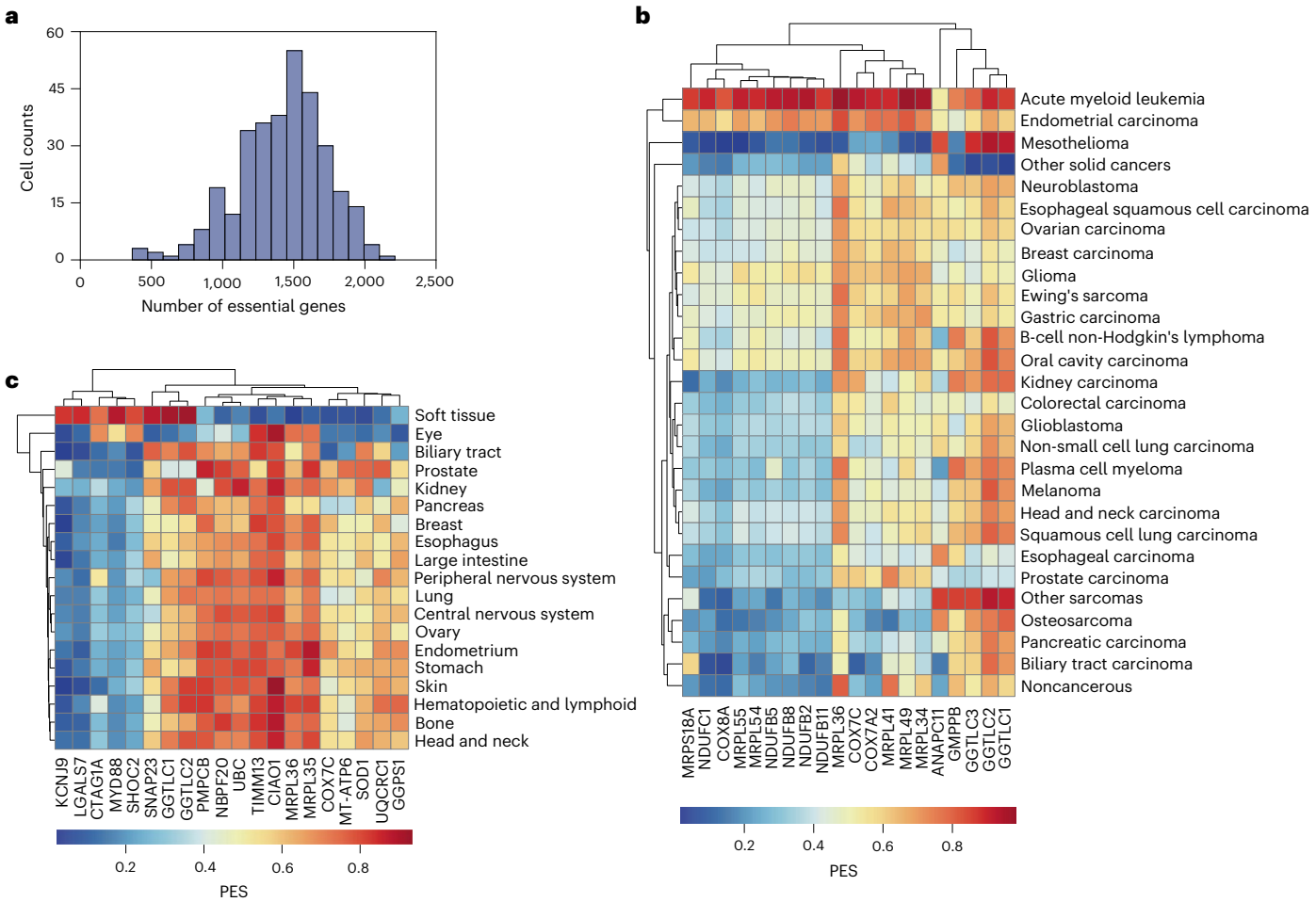


Fig. 4 | Biological relevance of PES generated by PIC models. a, The Spearman correlation matrix between different levels of PES scores and other biological metrics. **b**, The trend boxplots between PES and six well-known biological metrics. **c**, The distribution of PES across proteins with distinct subcellular localizations ($n = 9{,}981$). **d**, BP enrichment analysis results of proteins with different PES. Proteins ranked by PES were divided into ten equal-sized group. **e**, KEGG enrichment analysis of ten protein bins with different PES. Group: proteins ranked by PES were divided into ten equal-sized group. Degree: protein node degree in the PPI network ($n = 18{,}726$). PhastCons: conservation score across species ($n = 16{,}437$). PhyloP: phylogenetic *P* value indicating evolutionary conservation ($n = 16{,}437$). NofDiseases: number of diseases related to a certain protein ($n = 15{,}190$). Normal-expr: protein expression level in normal cells ($n = 19{,}363$). Cancer-expr: protein expression level in cancer cells ($n = 19{,}363$). *n* represents the number of samples used to generate each subplot. The box plots show the median (center line), the 25th and 75th percentiles (bounds of the box) and the minimum and maximum values (whiskers). The orange line in each box plot represents a linear fit. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. *P* values were calculated using Fisher's exact test (one-sided, $n = 18{,}870$, where *n* represents the total number of genes).

strong correlation ($\rho = 0.83$, $P < 1.0 \times 10^{-323}$), whereas hPES shows weaker correlations with mPES ($\rho = 0.26$, $P = 5.0 \times 10^{-291}$) and cPES ($\rho = 0.25$, $P = 5.5 \times 10^{-290}$) (Fig. 4a and Supplementary Table 7). These findings align with the observed divergence in protein essentiality across three levels, underscoring the importance of assessing protein essentiality comprehensively via different levels of PES.

Furthermore, we analyzed the variations in human protein essentiality across 28 types of cancer and 19 types of tissue. Specifically,

https://doi.org/10.1038/s43588-024-00733-1



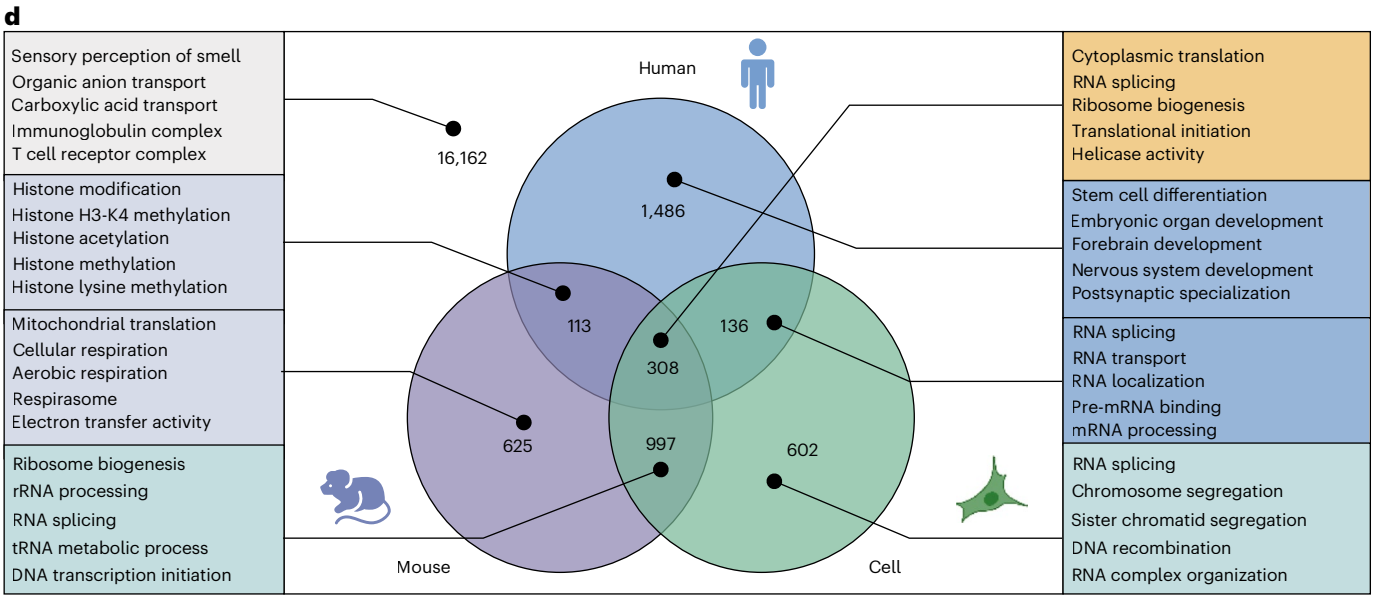


Fig. 5 | Cross-level analyses based on PES at different levels. a, The distribution of the number of essential proteins across 323 human cell lines. **b**, Top 20 proteins with the highest variation in PES across different cancers. **c**, Top 20 proteins with the highest variation in PES across different tissues. **d**, Functional enrichment analysis results of DEPs across three levels. The numbers in the figure (central) represent the count of DEPs. The descriptions in the figure (side) denote the functional annotations corresponding to these DEPs.

we first averaged the PES generated by cell-level PIC models sourced from the same disease or tissue to measure protein essentiality in that disease or tissue. Then, we calculated the standard deviation of PES for all human proteins across diseases and tissues, and selected the top 20 proteins with the highest standard deviations. As a result, we observed substantial differences in protein essentiality between nonsolid tumors (for example, acute myeloid leukemia) and solid tumors (for example, breast carcinoma) (Fig. 5b). Functional annotation revealed that these proteins (for example, NDUFC1), which exhibit higher essentiality in nonsolid tumors but lower essentiality in solid tumors, are typically involved in oxidative phosphorylation (OXPHOS). OXPHOS, as a part of the aerobic respiration, is commonly deficient in solid tumors because the tumor microenvironment in solid tumors is highly hypoxic, leading solid tumor cells to rely predominantly on anaerobic respiration or glycolysis for energy generation. Consequently, these proteins involved in OXPHOS are essential in nonsolid tumors but nonessential in solid tumors. A few current studies have also supported our findings[25,26]. Moreover, we also found that some proteins (for example, LGALS7) are essential in soft tissue but nonessential in other tissues (Fig. 5c). Functional annotation via the UniProt[27] database revealed that these proteins are involved primarily in cell–cell and cell–matrix interactions, which are commonly seen in the soft tissue.

Furthermore, we tried to discover the functional divergence of essential proteins across three levels via PES. First, we defined differentially essential proteins (DEPs) as the proteins that have different labels predicted by PIC at the human, mouse and cell level. Proteins ranked in the top 10% sorted by hPES, mPES and cPES are considered essential at each level, while the rest of the proteins are considered nonessential. Therefore, eight groups of DEPs were constructed by combining predicted labels from the three levels ($2 \times 2 \times 2 = 8$). Subsequently, Gene Ontology and KEGG enrichment analyses were performed on the DEPs in the eight groups to obtain related function annotation terms. The results showed that DEPs in different groups perform diverse functions (Fig. 5d and Supplementary Data 4 and 5). For example, it is apparent that proteins marked essential only in humans play a key role in maintaining organogenesis and development in complex systems, especially in the nervous system. In contrast, proteins marked essential in both mouse and human cell lines are involved in genetic information transmission and providing energy for cell proliferation. Given that most of the cells in humans are highly differentiated with low proliferative capacity while those in mouse embryos and human cell lines have high proliferative activity, it is no wonder that the essential proteins identified at different levels carry out diverse functions. Therefore, this further suggested that, for humans, human cell lines and mice, the difference in cellular physiological stage results in the divergence in functional requirements and then leads to the divergence in protein essentiality.

Case studies

Breast cancer is the most common malignancy among women, posing a formidable health challenge worldwide[28]. Therefore, it is important to develop biomarkers that can efficiently predict the prognosis of breast cancer. To test the potential of the proposed PES algorithm in this task, we collected breast invasive carcinoma (BRCA) project data from the Cancer Genome Atlas (TCGA) database. Specifically, we downloaded patients' transcriptome data and clinical information from TCGA. After filtering out genes with low-level expression and patient data with incomplete clinical survival information, we ultimately obtained the expression profile data of 19,938 genes for 1,115 patients with breast cancer. Using an ensemble learning strategy, we integrated the 25 PIC models trained on breast cancer cell lines into a general breast cancer-level PIC model, denoted as PIC-BRCA. Furthermore, by using PIC-BRCA, we calculated the PES in breast cancer, denoted as PES-BRCA. Proteins with higher PES-BRCA are expected to be more essential for the survival of breast cancer cells, which then could serve as potential therapeutic targets or prognostic biomarkers. We ranked all proteins by PES-BRCA and then selected the top ten proteins with the highest PES. Subsequently, we categorized the 1,115 patients with breast cancer into high-expression and low-expression groups based on the median expression level of the ten proteins in the patients. log-rank tests were performed between the high-expression and the low-expression groups to investigate whether the ten proteins are associated with the survival of breast cancer. All ten proteins are able to predict the survival of breast cancer (log-rank test, $P < 0.01$). In addition, six out of the ten proteins have been supported by publications[29–34] (Supplementary Table 8). This case study indicates that PES is able to discover prognostic biomarkers or even therapeutic targets. It is thus expected that PES could assist in finding prognostic biomarkers and therapeutic targets in other types of cancer and disease because we can combine various cell-level PIC models to measure protein essentiality as needed.

To further experimentally validate the ten proteins we have screened out, we additionally collected six ground-truth clinical breast cancer patient cohorts from the Gene Expression Omnibus (GEO) database to verify that these ten proteins can serve as new prognostic markers in these additional cohorts (Fig. 6a). We conducted the experimental validation following the same pipeline as above. Experimental results showed that eight out of these ten proteins can serve as prognostic biomarkers for breast cancer in more than half of the cohorts (Fig. 6b and Supplementary Data 6). In particular, proteasome 26S subunit, non-ATPase 7 (PSMD7) was validated as an effective prognostic biomarker in almost all cohorts, demonstrating its robust potential as a prognostic biomarker for breast cancer (Fig. 6c).

In the human genome, there are approximately 20,000 protein-coding RNAs, but there are millions of microproteins translated from small open reading frames (smORFs), which are open reading frames with lengths within 100 codons[35]. Existing literature has confirmed that microproteins play crucial roles in regulating cell proliferation[36], cellular respiration[37] and immune modulation[38]. However, there are still no methods available for measuring the essentiality of microproteins up to now. It is thus important to test whether PES is able to evaluate human microprotein essentiality or not. Ji et al. proposed smORFunction, a tool that integrates 617,462 unique smORFs and their corresponding microproteins, for the functional prediction of these microproteins[35]. We obtained the sequences of all human microproteins from smORFunction and used the PIC models to predict the essentiality of the 617,462 microproteins by calculating their corresponding PES[39]. To confirm the efficacy of PES in measuring microprotein essentiality, we selected the top ten microproteins with the highest PES (Supplementary Table 9) and then used smORFunction to predict the functions of the ten microproteins. The results showed that these microproteins are involved mainly in essential biological processes such as cell division, cellular respiration and DNA replication (Supplementary Data 8), which is consistent with the results for essential proteins. These results suggest that PES could be also used to efficiently quantify the essentiality of microproteins.

Discussion

Essential proteins, encoded by essentials genes, are crucial for an organism's survival, often participating in fundamental biological processes. Identifying essential proteins from the human proteome is therefore of great importance for disease prevention, diagnosis and treatment. However, no single protein is absolutely essential; only functions can be so[23]. Human protein essentiality is context dependent and closely associated with cell type and physiological stage. Furthermore, human essential proteins differ greatly across living humans, human cell lines and animal models. Our PIC model considers the variation in human protein essentiality across different levels, predicting essential scores for proteins in living humans, human cell lines and animal models at the same time. We believe that PIC will be beneficial

https://doi.org/10.1038/s43588-024-00733-1

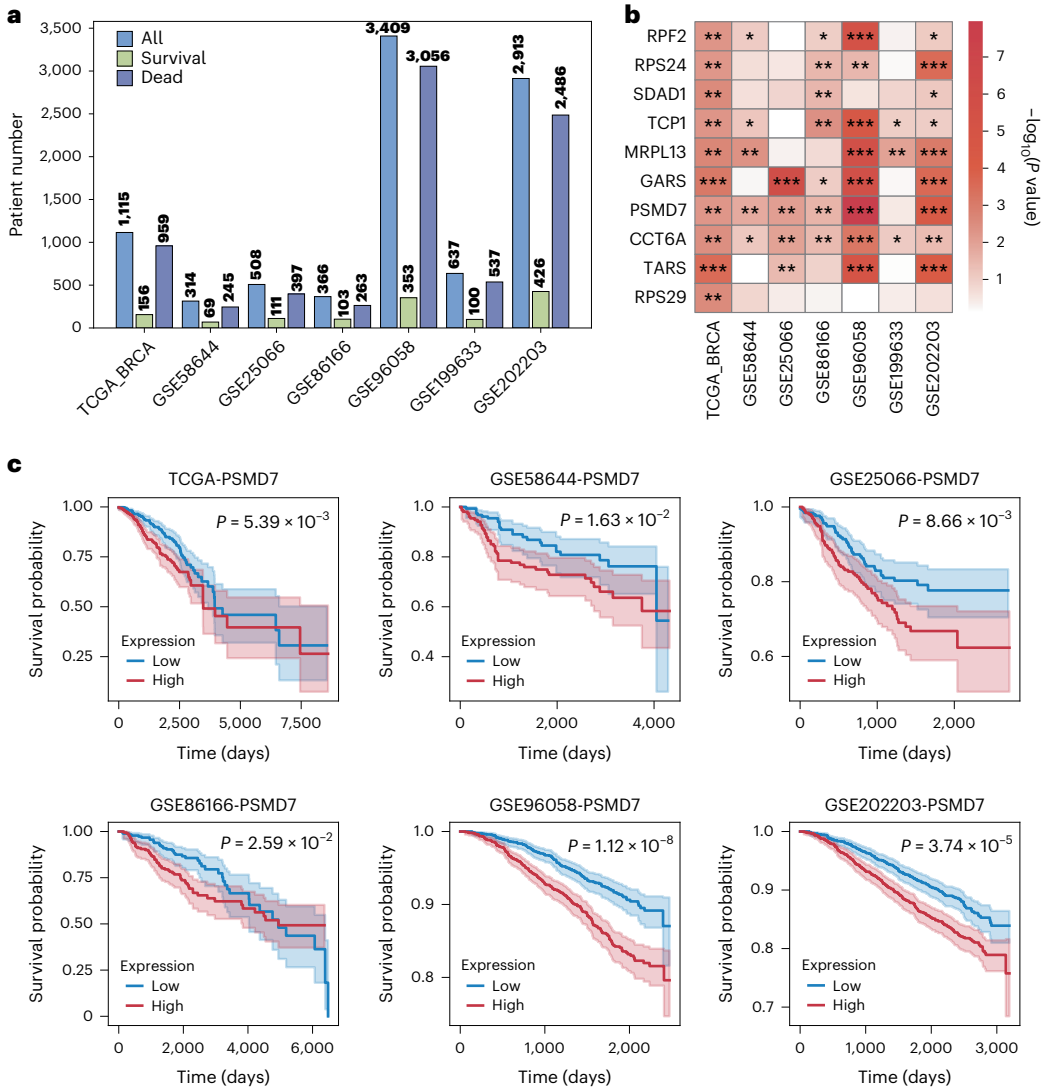


Fig. 6 | Finding potential prognostic biomarkers for breast cancer through PES. a, The distribution of patient numbers across different breast cancer cohorts from the TCGA and GEO databases. **b**, The prognostic value of ten proteins selected using PES across seven clinical breast cancer patient cohorts. Each cell in the heatmap represents the $-\log_{10}(P\text{ value})$ of the protein's association with patient survival, as determined by log-rank tests. The columns represent different datasets from the TCGA and GEO databases, while the rows represent the proteins analyzed. **c**, Kaplan–Meier survival curves for PSMD7. Patients were divided into high- and low-expression groups based on the median expression level of PSMD7 within cohorts. The shaded areas around each line represent the 95% confidence interval. *P* values were calculated using the log-rank test (two-sided, where the sample size *n* is the number of patients in each cohort). $^{*}P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$.

for users to comprehensively predict and understand the essentiality of human proteins, aiding in the discovery of therapeutic targets and prognostic biomarkers.

Future explorations and improvements in this domain include the following. (1) Enhancing the interpretability of predicted PES. Although we defined the PES using the probability values outputted by PIC and conducted preliminary exploration and analysis of its biological meaning, we lack in-depth explanation of the core biological meaning of PES, which is largely due to the fact that the neural network model is a black box. (2) Predicting and investigating protein essentiality across different species. Currently, PIC is limited to predicting human protein essentiality at three levels and lacks the capability to predict essentiality in other species, such as bacteria or other microorganisms. This constraint is mainly due to the scarcity of essentiality data for nonhuman species. In the future, constructing either a unified model or multiple species-specific models to predict protein essentiality could allow the investigation of cross-species commonalities and differences in essential proteins. This could have important implications for drug discovery, such as targeting essential proteins in bacteria for antibiotic development. (3) Incorporating protein structural information to enhance prediction performance. PIC model is a sequence-based deep learning model that predicts protein essentiality solely on the basis of the input protein sequence. However, the absence of structural information may limit the model's performance. Future models could integrate protein structural features, potentially leading to more accurate predictions of protein essentiality.

Methods

Data collection

We aimed to collect as much protein essentiality data as possible, with sample sizes at each level determined by the number of experimentally validated proteins available in the respective public databases. We did not use statistical methods to predetermine the sample size.

In addition, we did not exclude any protein essentiality data, as each protein's essentiality holds potential research value for biomedical studies.

For the human level, we obtained all human gene transcripts and their corresponding essentiality metric from the gnomAD database (V4.0.0). Loss of function observed/expected upper bound fraction (LOEUF) is a type of gene constraint metric for measuring gene essentiality at the human level, which was calculated using data from large-scale sequencing projects conducted within the human population, representing the ratio of observed loss-of-function variants to the expected number of mutants in each gene transcript. The reason why LOEUF could reflect human gene essentiality is that essential genes/proteins typically exhibit intolerance to loss-of-function variants. Consequently, the number of observed loss-of-function variants in essential gene transcripts within healthy human cohorts is expected to be substantially lower than the expected number. This implies that the LOEUF value of essential gene transcripts should be correspondingly lower[1]. Here, we used the officially recommended LOEUF threshold as the criterion for essentiality classification. Proteins with LOEUF values below 0.6 were categorized as essential, while those with LOEUF values equal to or above 0.6 were classified as nonessential. We preprocessed the raw data obtained from gnomAD by removing missing and redundant values and then used Ensembl BioMart[40] to align the transcript IDs of genes with their respective protein sequences. As a result, we obtained 65,057 protein sequences derived from 17,552 protein-coding genes, along with their corresponding LOEUF values, including 14,146 positive samples and 50,911 negative samples.

For the mouse level, given that mouse serves as a popular model organism in medicine[1], it is important to investigate the essentiality of human–mouse homologous proteins. To achieve this, we used homologous human protein sequences as inputs for the PIC-mouse model and assigned essentiality labels based on the essentiality data of homologous mouse proteins. We retrieved mouse gene essentiality data from the OGEE database, which is sourced from the Mouse Genome Informatics (MGI) database. Subsequently, we used Ensembl BioMart to match the gene IDs with the corresponding protein sequences. This process generated 6,050 human protein sequences with their homologous mouse protein essentiality label, comprising 443 positive samples and 5,607 negative samples.

For cell line level, we collected the human cell line-specific binary essential score matrix from the Project Score database, which were generated by Wellcome Sanger Institutes using the systematic CRISPR–Cas9 knockout screening technique in diverse human cell lines. This matrix contains varying binary essential scores for 17,995 human protein-coding genes across 323 human cell lines. Then, we used Ensembl BioMart to align the gene symbols with their corresponding protein sequences. Finally, we obtained the binary essentiality labels for 17,185 protein sequences across 323 human cell lines originated from 19 human tissues and 28 human diseases.

Model architecture

PIC-human, PIC-mouse and PIC-cell share identical model architectures. PIC model comprises three main modules: Embedding, Attention and Prediction (Fig. 1b).

The purpose of the Embedding module is to extract features from protein sequences. We used a pretrained PLM to convert variable-length raw protein sequences into fixed-size numerical feature vectors, which is the classical embedding process.

The embedding process is summarized as follows.

A protein sequence is typically represented as

$$p = \{aa_1, aa_2, \ldots, aa_i, \ldots, aa_L\}, \quad (1)$$

where p denotes a protein sequence, aa_i represents the amino acid residue at the ith position in the sequence and L is the length of the protein sequence.

Then, using the ESM-2 model, we encoded each amino acid residue in the protein sequence into a 1,280-dimensional numerical feature vector, denoted as $\mathbf{x}$.

Consequently, we obtained a residue-level feature matrix $X \in R^{L\times d}$ to represent the entire protein sequence:

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_L], \quad (2)$$

where L represents the length of the protein sequence and d represents the dimension of embedding features.

After the embedding module, we applied a multihead attention module to capture the relative importance of amino acids at various positions within protein sequences. The attention module can transform the initial sequence feature matrix $X \in R^{L\times d}$ generated by embedding the module into a new sequence feature matrix weighted by attention weights, denoted as $X' \in R^{L\times d}$. This new matrix incorporates information regarding the relative importance of amino acids at different positions in the sequences.

The multihead attention mechanism is formulated as follows:

$$\text{MultiHead}(Q,K,V) = \text{Norm}\left(\text{Concat}\left(\text{head}_1 \ldots, \text{head}_i, \ldots, \text{head}_n\right) W^o\right) \quad (3)$$

$$\text{head}_i = \text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (4)$$

where head_i denotes the ith attention head. Q, K and V represent query, key and value, respectively. $\frac{1}{\sqrt{d_k}}$ is the scaling factor of the dot-product attention. 'T' denotes the matrix transposition. 'Norm' refers to the layer normalization operation. 'Concat' represents the concatenation operation. W^o represents the weight matrix for output transformation.

In the Prediction module, we used an average-pooling layer to downsample the high-dimensional feature representation, which converts the residue-level feature $X' \in R^{L\times d}$ into the sequence-level feature $X'' \in R^{1\times d}$ and ensures a fixed dimensionality of the model input. Then, the sequence-level feature X'' was fed into a multilayer perceptron to generate the raw prediction score. Finally, the raw prediction score was passed through the sigmoid activation function to obtain the predicted essential probability of the protein sequence.

Definition of the PES

Existing computational methods focus primarily on the classification of essential proteins, which overlooked the exploration and analysis of the biological relevance of the probability values output by a predictive model. The probability values generated by PIC represent the likelihood that a given protein is essential, which could serve as a potential quantitative metric for estimating protein essentiality. Here, we defined the probability values output by the PIC-human, PIC-mouse and PIC-cell models as the essential scores of human proteins at the human level, mouse level and cell level, denoted as hPES, mPES and cPES, respectively. To establish a comprehensive metric for evaluating human protein essentiality, we then calculated the mean of hPES, mPES and cPES as the PES for each protein, which could measure protein essentiality across three levels. We hypothesized that PES can more comprehensively reflect human protein essentiality because it integrates probability values from models at three different levels. We explored the biological importance of PES and demonstrated that it could serve as a comprehensive metric for measuring human protein essentiality. We also calculated the mean of probability values output by cell-level PIC models originating from the same tissue or disease, thereby obtaining PES at the tissue level and the disease level.

Web server

We have developed a user-friendly web server that is freely accessible on the PIC website (http://www.cuilab.cn/pic). This tool allows users to

input the candidate protein sequences and then obtain the essentiality of the input proteins at three levels. The results can be downloaded from the result page. The website was constructed on the basis of the packages of Python 3, Flask and Numpy.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Protein essentiality data were collected from gnomAD (https://gnomad.broadinstitute.org/), Project Score (https://www.sanger.ac.uk/tool/project-score-database/) and OGEE (https://v3.ogee.info/#/home) databases to train PIC-human, PIC-cell and PIC-mouse, respectively. PPI network node degree is accessible through the STRING (https://cn.string-db.org/) database. Gene expression values in normal tissue and cancer tissue are accessible through the Human Protein Atlas (https://www.proteinatlas.org/) database. Phylop and phastCons are accessible through the UCSC genome browser (https://genome.ucsc.edu/). The numbers of protein-related diseases are accessible through the DisGeNet (https://www.disgenet.com/) database. The transcriptome data and clinical information for patients with breast cancer can be collected from the TCGA (https://www.cancer.gov/ccg/research/genome-sequencing/tcga) database. Additional experimental validation cohort data were obtained from the GEO (https://www.ncbi.nlm.nih.gov/geo/) database including GSE58644, GSE25066, GSE86166, GSE96058, GSE199633 and GSE202203. The protein sequences of microproteins are accessible through the smORFunction (http://www.cuilab.cn/smorfunction) website. The data used in this study are available on Zenodo at https://doi.org/10.5281/zenodo.13994480 (ref. 39). Source data are provided with this paper.

Code availability

The PIC web server is available at http://www.cuilab.cn/pic. The PIC source code is available on GitHub at https://github.com/KangBoming/PIC and Zenodo at https://doi.org/10.5281/zenodo.13994480 (ref. 39).

References

1. Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Human gene essentiality. *Nat. Rev. Genet.* **19**, 51–62 (2018).
2. Ji, X., Rajpal, D. K. & Freudenberg, J. M. The essentiality of drug targets: an analysis of current literature and genomic databases. *Drug Discov. Today* **24**, 544–550 (2019).
3. Aromolaran, O., Aromolaran, D., Isewon, I. & Oyelade, J. Machine learning approach to gene essentiality prediction: a review. *Brief. Bioinf.* **22**, bbab128 (2021).
4. Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* **2005**, 96–103 (2005).
5. Wuchty, S. & Stadler, P. F. Centers of complex networks. *J. Theor. Biol.* **223**, 45–53 (2003).
6. Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**, 803–806 (2005).
7. Li, G. et al. Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinf.* **17**, 279 (2016).
8. Guo, F. B. et al. Accurate prediction of human essential genes using only nucleotide composition and association information. *Bioinformatics* **33**, 1758–1764 (2017).
9. Hasan, M. A. & Lonardi, S. DeeplyEssential: a deep neural network for predicting essential genes in microbes. *BMC Bioinf.* **21**, 367 (2020).
10. Zhang, X., Xiao, W. & Xiao, W. DeepHE: accurately predicting human essential genes based on deep learning. *PLoS Comput. Biol.* **16**, e1008229 (2020).
11. Zeng, M. et al. A deep learning framework for identifying essential proteins by integrating multiple types of biological information. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **18**, 296–305 (2021).
12. Li, Y., Zeng, M., Wu, Y., Li, Y. & Li, M. Accurate prediction of human essential proteins using ensemble deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **19**, 3263–3271 (2022).
13. Li, Y., Zeng, M., Zhang, F., Wu, F. X. & Li, M. DeepCellEss: cell line-specific essential protein prediction with attention-based interpretable deep learning. *Bioinformatics* **39**, btac779 (2023).
14. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
15. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
16. Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H. & Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* **50**, W228–W234 (2022).
17. Teufel, F. et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
18. Hou, X., Wang, Y., Bu, D., Wang, Y. & Sun, S. EMNGly: predicting N-linked glycosylation sites using the language models for feature extraction. *Bioinformatics* **39**, btad650 (2023).
19. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
20. Eppig, J. T. Mouse Genome Informatics (MGI) Resource: genetic, genomic, and biological knowledgebase for the laboratory mouse. *ILAR J.* **58**, 17–41 (2017).
21. Dwane, L. et al. Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. *Nucleic Acids Res.* **49**, D1365–D1372 (2021).
22. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
23. Chen, H. et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. *Brief. Bioinf.* **21**, 1397–1410 (2020).
24. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
25. Wang, J. D., Xu, J. Q., Long, Z. J. & Weng, J. Y. Disruption of mitochondrial oxidative phosphorylation by chidamide eradicates leukemic cells in AML. *Clin. Transl. Oncol.* **25**, 1805–1820 (2023).
26. Liu, L. et al. High metabolic dependence on oxidative phosphorylation drives sensitivity to metformin treatment in MLL/AF9 acute myeloid leukemia. *Cancers* **14**, 486 (2022).
27. UniProt Consortium. UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
28. Jabbarzadeh Kaboli, P. et al. Unlocking c-MET: a comprehensive journey into targeted therapies for breast cancer. *Cancer Lett.* **588**, 216780 (2024).
29. Zheng, L. et al. A potential tumor marker: chaperonin containing TCP-1 controls the development of malignant tumors (Review). *Int. J. Oncol.* **63**, 106 (2023).
30. Cai, M., Li, H., Chen, R. & Zhou, X. MRPL13 promotes tumor cell proliferation, migration and EMT process in breast cancer through the PI3K–AKT–mTOR pathway. *Cancer Manag. Res.* **13**, 2009–2024 (2021).
31. Zhao, Y. et al. Deubiquitinase PSMD7 regulates cell fate and is associated with disease progression in breast cancer. *Am. J. Transl. Res.* **12**, 5433–5448 (2020).
32. Vishnubalaji, R. & Alajez, N. M. Single-cell transcriptome analysis revealed heterogeneity and identified novel therapeutic targets for breast cancer subtypes. *Cells* **12**, 1182 (2023).
33. Gui, Z., Liu, P., Zhang, D. & Wang, W. Clinical implications and immune implications features of TARS1 in breast cancer. *Front. Oncol.* **13**, 1207867 (2023).

34. Song, S. et al. CHMP4A stimulates $CD8^+$ T-lymphocyte infiltration and inhibits breast tumor growth via the LSD1/IFNβ axis. *Cancer Sci.* **114**, 3162–3175 (2023).
35. Ji, X., Cui, C. & Cui, Q. smORFunction: a tool for predicting functions of small open reading frames and microproteins. *BMC Bioinf.* **21**, 455 (2020).
36. Polycarpou-Schwarz, M. et al. The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* **37**, 4750–4768 (2018).
37. Makarewich, C. A. et al. MOXI is a mitochondrial micropeptide that enhances fatty acid β-oxidation. *Cell Rep.* **23**, 3701–3709 (2018).
38. Bhatta, A. et al. A mitochondrial micropeptide is required for activation of the Nlrp3 inflammasome. *J. Immunol.* **204**, 428–437 (2020).
39. Kang, B., Fan, R., Cui, C. & Cui, Q. Comprehensive prediction and analysis of human protein essentiality based on a pre-trained protein large language model(v1.0). *Zenodo* https://doi.org/10.5281/zenodo.13994480 (2024).
40. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).

Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (62025102, 32301239 and 81921001) and the Scientific and Technological Research Project of Xinjiang Production and Construction Corps (2023AB057).

Author contributions

Q.C. presented the original idea. B.K. and R.F. designed the study. B.K. performed the study. B.K., R.F., C.C. and Q.C. wrote or edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-024-00733-1.

Correspondence and requests for materials should be addressed to Qinghua Cui.

Peer review information *Nature Computational Science* thanks Min Li and Stefano Lonardi for their contribution to the peer review of this work. Primary Handling Editor: Jie Pan, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Qinghua Cui

Last updated by author(s): Oct 26, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	We built and used PIC for downstream data analysis. The source code developed for this study is available at GitHub (https://github.com/KangBoming/PIC) and Zenodo (https://zenodo.org/records/13994480). We employed Pandas (v2.1.1), Numpy (v1.2.6) and Scipy (v1.14.1) to process the data and perform statistics tests. We used scikit-learn (v1.3.2) to randomly split the data into training, validation, and test sets.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used in this study were obtained from publicly available datasets or online websites. Specifically, protein essentiality data were collected from gnomAD (https://gnomad.broadinstitute.org/), Project Score (https://www.sanger.ac.uk/tool/project-score-database/) and OGEE (https://v3.ogee.info/#/home) databases to

train PIC-human, PIC-cell and PIC-mouse respectively. Protein-protein interaction network node degree could be accessible through the STRING (https://cn.string-db.org/) database. Gene expression value in normal tissue and cancer tissue could be accessible through the HPA (https://www.proteinatlas.org/) database. Phylop and phastCons could be accessible through the UCSC genome browser (https://genome.ucsc.edu/). The numbers of protein-related diseases could be accessible through the DisGeNet (https://www.disgenet.com/) database. The transcriptome data and clinical information for breast cancer patients could be collected from the TCGA (https://www.cancer.gov/ccg/research/genome-sequencing/tcga) database. Additional experimental validation cohort data were obtained from the GEO (https://www.ncbi.nlm.nih.gov/geo/) database including GSE58644, GSE25066, GSE86166, GSE96058, GSE199633, and GSE202203.The protein sequences of microproteins could be accessible through the smORFunction (http://www.cuilab.cn/smorfunction) website. Source data for Figures 2,3,4,5,6 are available with this manuscript.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	To comprehensively train and evaluate the PIC models, we collected protein essentiality data at the human,cell-line and mouse levels. The sample size of protein essentiality data at each level was determined by the number of experimentally validated entires available in the corresponding databases. Consequently, we obtained a total of 65057 samples for training PIC-human model, 17185 samples for training PIC-cell model, and 6050 samples for training PIC-mouse model.
Data exclusions	No data were excluded from the analyses.
Replication	We set random seeds for data splitting and model training in order to reproduce the experimental findings.
Randomization	Protein essentiality data were randomly split into training, validation, and test sets using scikit-learn, ensuring distinct data in each set.
Blinding	Blinding is irrelevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A