Cat-E: A comprehensive web tool for exploring cancer targeting strategies

Rana Salihoglu [a], Johannes Balkenhol [a,b], Gudrun Dandekar [c], Chunguang Liang [a,d], Thomas Dandekar [a,e,**], Elena Bencurova [a,*]

[a] Department of Bioinformatics, University of Wurzburg, 97074 Wurzburg, Germany
[b] Rudolf Virchow Center for Integrative and Translational Bioimaging, University Hospital of Wurzburg, 97080 Wurzburg, Germany
[c] Chair of Tissue Engineering and Regenerative Medicine, University Hospital of Wurzburg, 97080 Wurzburg, Germany
[d] Institute of Immunology, Jena University Hospital, Friedrich-Schiller-University, 07743 Jena, Germany
[e] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

ARTICLE INFO

ABSTRACT

Identifying potential cancer-associated genes and drug targets from omics data is challenging due to its diverse sources and analyses, requiring advanced skills and large amounts of time. To facilitate such analysis, we developed Cat-E (**Ca**ncer **T**arget **E**xplorer), a novel R/Shiny web tool designed for comprehensive analysis with evaluation according to cancer-related omics data. Cat-E is accessible at https://cat-e.bioinfo-wuerz.eu/. Cat-E compiles information on oncolytic viruses, cell lines, gene markers, and clinical studies by integrating molecular datasets from key databases such as OvirusTB, TCGA, DrugBANK, and PubChem. Users can use all datasets and upload their data to perform multiple analyses, such as differential gene expression analysis, metabolic pathway exploration, metabolic flux analysis, GO and KEGG enrichment analysis, survival analysis, immune signature analysis, single nucleotide variation analysis, dynamic analysis of gene expression changes and gene regulatory network changes, and protein structure prediction. Cancer target evaluation by Cat-E is demonstrated here on lung adenocarcinoma (LUAD) datasets. By offering a user-friendly interface and detailed user manual, Cat-E eliminates the need for advanced computational expertise, making it accessible to experimental biologists, undergraduate and graduate students, and oncology clinicians. It serves as a valuable tool for investigating genetic variations across diverse cancer types, facilitating the identification of novel diagnostic markers and potential therapeutic targets.

1. Introduction

Cancer is a common disease that significantly threatens human life. According to the World Health Organization (WHO), cancer causes around 9.6 million deaths annually, or around one in six fatalities worldwide [1]. Modern medical research is rapidly expanding, providing new opportunities driven by increasing amounts of cancer-related data. The challenge uniquely answered by Cat-E is to evaluate molecular targets and innovative protein targeting strategies in the light of the massive volume of available omics data. Notable resources such as Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) have significantly enhanced data analysis and the ability to understand gene functions (TCGA: https://tcga-data.nci.nih.gov/tcga; GTEx:https://gtexportal.org/home/). Access to these resources has led to the development of numerous pipelines for

investigating large gene expression datasets that are now used as standards in data analysis. For example, survival analysis is essential for examining the relationships between gene expression levels and prognostic outcomes, particularly in evaluating the clinical significance of specific drug targets assessed. Several web tools such as TSVdb [2], Gepia2 [3], and cBioPortal [4] have been developed to analyze differential gene expression (DGE) using TCGA data. For ease of use, Cat-E was developed as a one-stop R/Shiny application integrating routines for diverse omics data and features, including expression and protein structure, for rapid cancer drug target evaluation.

R/Shiny applications have become increasingly popular in bioinformatics, advancing the interpretation and analysis of biological data. Popular tools are for example GENAVi [5], which allows the user to load and analyze enriched and DGE data; ShinyGO [6] and PRO-Simat [7], allowing both dynamic simulations and pathway enrichment analysis;

---

and GEOexplorer web server [8], conducting DGE and enrichment analyses with GEO datasets.

The study of gene mutations in cancer research has significantly improved using TCGA data. Tools such as Cancer3D [9] and IntOGen [10] are crucial for locating and describing genetic changes linked to various types of cancer. MIMOSA2 [11] facilitates the modelling and evaluation of relationships between microbiota members and their metabolic products. Shiny GATOM [12] is an example of a Shiny web application designed for metabolic analysis.

Despite those tools offering outstanding contributions in cancer research, there are several limitations which we tackle here: The first is the lack of integration with diverse data sources, which might lead to incomplete analysis or even misinterpretation of the analysis. Moreover, many tools have limited scalability or computing resources, which can limit or slow down the analysis of large datasets. Furthermore, we found several such tools were not user-friendly and had nonintuitive interfaces, which required advanced technical or programming skills.

Our new, interactive web tool, **Ca**ncer **T**arget **E**xplorer (Cat-E), integrates published databases, datasets, and multiple analyses in a unique framework and toolkit, allowing evaluation of potential cancer-drug targets in the light of transcriptome and proteome data; this includes pathway analysis, metabolic flux analysis, immune signature analysis, survival analysis, and single nucleotide variation analysis. We designed Cat-E to facilitate research of new cancer targets and protein structures as well as new cancer targeting strategies related to cancer pathways and weak immune responses. The intended users are experimental biologists, undergraduate and graduate students, and oncology research-minded clinical scientists. Our goal is to simplify access to cancer-

Fig. 1. Schematic representation of Cat-E web tool analyses workflow. Cat-E implements nine functional analyses related to cancer and oncolytic viruses. Blue arrows depict the logical flow of an analysis.

related transcriptome and omics data, offer a central data warehouse with detailed and trustworthy information, and enable users to conduct bioinformatic analyses efficiently, even without programming skills. We are confident that Cat-E is a welcome tool for evaluating various anti-cancer targets (mainly proteins, and with suitable omics data, also genes and RNAs) in different types of cancer including evaluation in the light of prognostic markers.

2. Materials and methods

Cat-E was implemented using Shiny package of the R programming language (version 4.2.3), and MariaDB (https://mariadb.org). As explained in the following subsections, we incorporated the Python and Java scripts and multiple R packages to implement various analyses (Fig. 1).

2.1. Data collection

Data on cell lines infected by oncolytic viruses were retrieved from the OvirusTDB database [14]. The following oncolytic viruses are available for analysis in Cat-E: Adenovirus, Herpes simplex virus, Measles virus, Canine parvovirus, Vaccinia virus, Alphavirus, Reovirus, Avian influenza A virus, Sindbis virus, Vesicular stomatitis virus, Parvovirus, Newcastle disease virus, Enterovirus, Respiratory syncytial virus, Mumps virus, Measles virus, Tanapoxvirus, Poxvirus, Sendai virus, Semliki forest virus, Maraba virus, Myxoma virus, and Bovine herpesvirus.

Moreover, diverse datasets on oncolytic virus strains were obtained from STRING viruses, intAct, and BioGRID [15–17], and a new database including these data was created using MariaDB. Datasets were obtained for human (*Homo sapiens*) and following viruses: Human adenovirus A serotype 12 (HadV-12), Human adenovirus A serotype 31, Human adenovirus B serotype 3, Human adenovirus C serotype 2 (HadV-2), Human adenovirus C serotype 5 (HadV-5), Human adenovirus D serotype 9, Human adenovirus E, Human adenovirus 36, Human herpesvirus 1, Human herpesvirus 1 (strain 17), Human herpesvirus 8, Human herpesvirus 8 (HHV-8), Human herpesvirus 8 type M, Human herpesvirus, Murine adenovirus A serotype 1 (MadV-1), Newcastle disease virus, Enterovirus, Respiratory syncytial virus, Mumps virus, Measles virus, Tanapoxvirus, Poxvirus, Reovirus type 1 (strain Lang) (T1L), Reovirus type 3 (strain Dearing), Sendai virus, Semliki forest virus, Sindbis virus, Vesicular stomatitis virus, Myxoma virus, Bovine herpesvirus, Avian orthoreovirus, Vaccinia virus, Vaccinia virus Copenhagen, Vaccinia virus WR, Vaccinia virus L-IPV, Vaccinia virus Ankara, Vaccinia virus GLV-1h68, and Vaccinia virus VVΔTKΔN1L.

The Cancer Genome Atlas (TCGA) and GTEx data were acquired using the TCGAbiolinks package [18], allowing differential expression and clinical analyses of 33 various tumor types.

The following tumor types are included in Cat-E: Adrenocortical carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Colon adenocarcinoma (COAD), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Esophageal carcinoma (ESCA), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Acute Myeloid Leukemia (LAML), Brain Lower Grade Glioma (LGG), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Mesothelioma (MESO), Ovarian serous cystadenocarcinoma (OV), Pancreatic adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate adenocarcinoma (PRAD), Rectum adenocarcinoma (READ), Sarcoma (SARC), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Testicular Germ Cell Tumors (TGCT), Thyroid carcinoma (THCA), Thymoma (THYM), Uterine Corpus Endometrial Carcinoma (UCEC),

Uterine Carcinosarcoma (UCS), and Uveal Melanoma (UVM).

Cell line information was collected from the DepMap database (https://depmap.org/portal/), including detailed tissue-specific cell line characteristics (e.g., cell line models, mutations, and DNA alterations). Furthermore, gene marker information was retrieved from the Cell Markers database [19].

To enhance the database embedded within Cat-E, we incorporated multiple data sources, including rPanglaoDB [20] for cell type information, DrugBANK [21] for drug descriptions, PubChem [22] for molecular details and drug structures, DGIdb [23] for drug-gene interactions, and STITCH [24] for drug-gene network information. Comprehensive details on clinical studies, including those on chimeric antigen receptor T-cell (CAR-T) treatment, bispecific antibody treatments, cytostatic therapies, and oncolytic virus therapies, were obtained from clinicaltrials.gov [25] and CKTTD checkpoint database [26].

2.2. The analytical capabilities of Cat-E

Cat-E provides a comprehensive approach to analyze cancer-related data, including gene expression, metabolic pathways, metabolic flux, GO and KEGG enrichment, survival, immune signature, single nucleotide variation, gene regulatory dynamics, and protein structure prediction (Fig. 1). It allows the user to upload their own data or select DGE data from the specified cancer types section to conduct a range of analyses. The details on the methodology and building of these analyses are described in the following subsections:

2.2.1. Differential gene expression (DGE) analysis
Datasets of more than 11,250 patient samples across 33 distinct cancer tissue types were collected by TCGA and GTEx databases through the TCGAbiolinks package [18]. Subsequently, data was normalized and filtered at the 25th quantile and DGE analysis was performed by limma [27], edgeR [28], and TCGAbiolinks [18]. Ensembl IDs were converted to Hugo Symbols using biomaRt package [29]. Cat-E provides filtered outcomes derived from Differential Gene Expression (DGE) analysis, presenting refined DGE data alongside tailored parameters, including log-fold change (log$_2$FC) thresholds, p-value thresholds, and options for selecting statistical methods such as limma or ANOVA. The resulting outcomes are visualized through tabular formats and graphical representations, including violin plots, volcano plots, and Principal Component Analysis (PCA) plots.

Cat-E also enables retrieval of GEO data using GSE identifiers through the GEOquery package [30], allowing limma-based DGE analysis by selecting log$_2$FC and p-value thresholds. It is also possible to upload the user's DGE analysis results and compare them with already published data using a Venn diagram, comparing common and unique genes.

2.2.2. Metabolic pathway analysis
The GATOM package and mwcsr package (Maximum Weight Connected Subgraph) [32] were introduced to analyze transcriptional and/or metabolic data to identify the most significant metabolic subnetwork [12]. Cat-E enables metabolic pathway analysis either by selecting 33 different cancer types or by uploading other datasets.

2.2.3. Metabolic flux analysis
Metabolic flux analysis is integrated by CNApy [31], a Python-based graphical user interface (GUI) derived from CellNetAnalyzer. This integration was achieved through a connection to R/Shiny, facilitated by the reticulate package [33]. SBML files of the metabolic network were created with the COBRA package [34] and visualized using d3flux (https://github.com/pstjohn/d3flux). These metabolic SBML/XML files enable Flux Balance Analysis (FBA), Parsimonious FBA, Flux Variability Analysis (FVA), and Elementary Flux Mode (EFM) analysis.

2.2.4. GO and KEGG enrichment analysis

R-package topGO and Kolmogorov-Smirnov (K-S) test were integrated to analyze GO terms [35], with both efficiency and sensitivity to topology. ClusterProfiler R package [36] was included to perform KEGG enrichment analysis for each of the 33 cancer types. Cat-E also allows modification of the protein-protein interactions (PPI) network and creation of the oncolytic virus models. KEGG analysis can be performed using the list of the human genes in the network with the parameters of "fdr", "maxGSsize = 500", and custom thresholds for p-value and q-value. Pathway representations are generated using log_2FC values obtained from the DGE data and are visualized with the help of the pathview package [37].

2.2.5. Survival analysis

Survival analysis evaluates the prognostic implications of various clinical factors. Clinical data were collected utilizing the TCGAbiolinks package [18] and survival analysis implemented via the survival [38] and survminer packages [39]. The results are visualized using the visNetwork [40], Plotly [41], and ggplot2 [42] packages, providing interactive and informative representations of the survival outcomes. This comprehensive approach facilitates a thorough examination of survival patterns and their association with relevant clinical variables.

2.2.6. Immune signature analysis

To provide insights into immunological aspects, imsig package [43] was integrated to analyze immune-related processes via selected collection of immune gene signatures obtained from a network-based deconvolution method. Specifically, imsig is dedicated to identification of gene signatures linked to different biological processes, including B cells, interferon, macrophages, monocytes, neutrophils, NK cells, and others. A conservative correlation threshold of r = 0.7 was used for feature selection and immune signature analysis to ensure robustness. Inter-gene correlations were calculated using the DiffCorr package [44], a computational tool created explicitly for evaluating differential correlation, and networks were visualized using the visNetwork package [40].

2.2.7. Single nucleotide variations (SNVs) analysis

SNVs analysis was implemented using genomic sequencing data obtained from TCGA. The information consisting of Mutation Annotation Format (MAF) files was obtained through TCGAbiolinks [18], providing a comprehensive overview of somatic mutations. The files contained vital data, including chromosome location, variant allele frequency, and functional consequences of the mutations.

The SNV analysis was conducted and an oncoplot generated using the Maftool R package [45] to display the mutation frequencies in the samples. The g3viz package [46] in Cat-E was employed to create lollipop diagrams, thereby enhancing interactivity and visualization of SNVs.

2.2.8. Dynamic gene regulatory analysis with Jimena

Jimena [47] was integrated into Cat-E, facilitating the simulation and analysis of dynamic gene regulatory networks. The database within Cat-E, comprising nearly 40,000 signaling networks obtained from OmnipathR [48], enables efficient gene searches and streamlines the creation of customized biological signaling pathways. Through the combination of Cat-E and Jimena, the manipulation and simulation of these networks can be conducted, providing valuable insights into the regulatory mechanisms of cellular processes. Moreover, this platform presents opportunities for drug discovery efforts by exploring potential therapeutic interventions via manipulation of signaling networks within a computer simulation.

2.2.9. Protein structure

Protein structure prediction plays a crucial role in elucidating the functional characteristics of proteins. AlphaFold data were retrieved from the AlphaFold DataBase [49],[50] using UniProt KnowledgeBase identifiers. Obtained data was processed and visualized in a three-dimensional format using the NGLVieweR package [51]. The visualization process involved color-coding of the protein structures based on their confidence scores (pLDDT), facilitating clear visualization and enabling comprehensive analysis of the predicted models. Specifically, the color scheme was as follows:

- Dark blue: Indicates a very high confidence level, where pLDDT > 90.
- Light blue: Represents a high confidence level, with 90 > pLDDT > 70.
- Yellow: Corresponds to a low confidence level, where 70 > pLDDT > 50.
- Orange: Indicates a very low confidence level, with pLDDT < 50.

Integration of 3Dmol.js into Cat-E enabled the generation of three-dimensional protein structures utilizing Protein Data Bank (PDB) codes [52]. This comprehensive approach ensured accurate representation and analysis of protein structures, providing valuable insights into their functional properties.

3. Results

This study introduces the Cat-E web tool for detailed analysis of gene expression, pathway enrichment, single nucleotide variations, survival patterns, immune signatures, metabolic flux, and structural characterization, particularly in cancer research. We identify significant targets for further research through DGE analysis and reveal essential cellular processes via GO enrichment. Examining individual changes in nucleotides and survival rates shows significant mutation clusters and distinct trends across various stages of the disease, respectively. Metabolic flux analysis identifies crucial pathways, such as gluconeogenesis, revealing possible targets for anticancer drug action, and AlphaFold can bring valuable information about protein structure. Taken together, this complex approach improves our comprehension of cancer biology and highlights the potential of Cat-E as a powerful tool for cancer research and therapeutic advancement.

We demonstrate the functionality of Cat-E using three distinct lung adenocarcinoma (LUAD) datasets (Fig. 2): DGE data available in Cat-E (use case 1), signaling network data from non-small cell lung carcinomas (NSCLC) [13] (use case 2), and glucogenesis pathway in cancer cells [53] (use case 3).

3.1. Use case 1: The analyses of LUAD data within the Cat-E database

The first use case of Cat-E is the LUAD dataset, which was chosen from among the 33 different types of cancer tissues available in Cat-E. The following section uses this dataset to show the results of several respective analyses: DGE, GO enrichment, SNV, survival, immune signature, and metabolic pathway exploration (Fig. 2).

3.1.1. DGE analysis

The DGE analysis on the LUAD dataset using the Limma method revealed 2146 genes identified as statistically significant ($log_2FC > 2$ and p-value < 0.001; Fig. 3A). Among these, 925 genes were upregulated, while 1221 genes displayed downregulation. The upregulated genes *golm1*, *pycr1*, *tedc2*, *tubb3*, and *muc16* emerged as compelling targets for further investigation. To demonstrate the comparison of DGE results, we used Cat-E to compare the LUAD data set with data from GEPIA2, using an interactive Venn diagram to analyze distinct and shared genes (n = 857) (as shown in Fig. 3B).

3.1.2. Gene ontology (GO) enrichment

Cat-E provides comprehensive features for conducting GO enrichment analysis. The findings are displayed using bar graphs and lists,
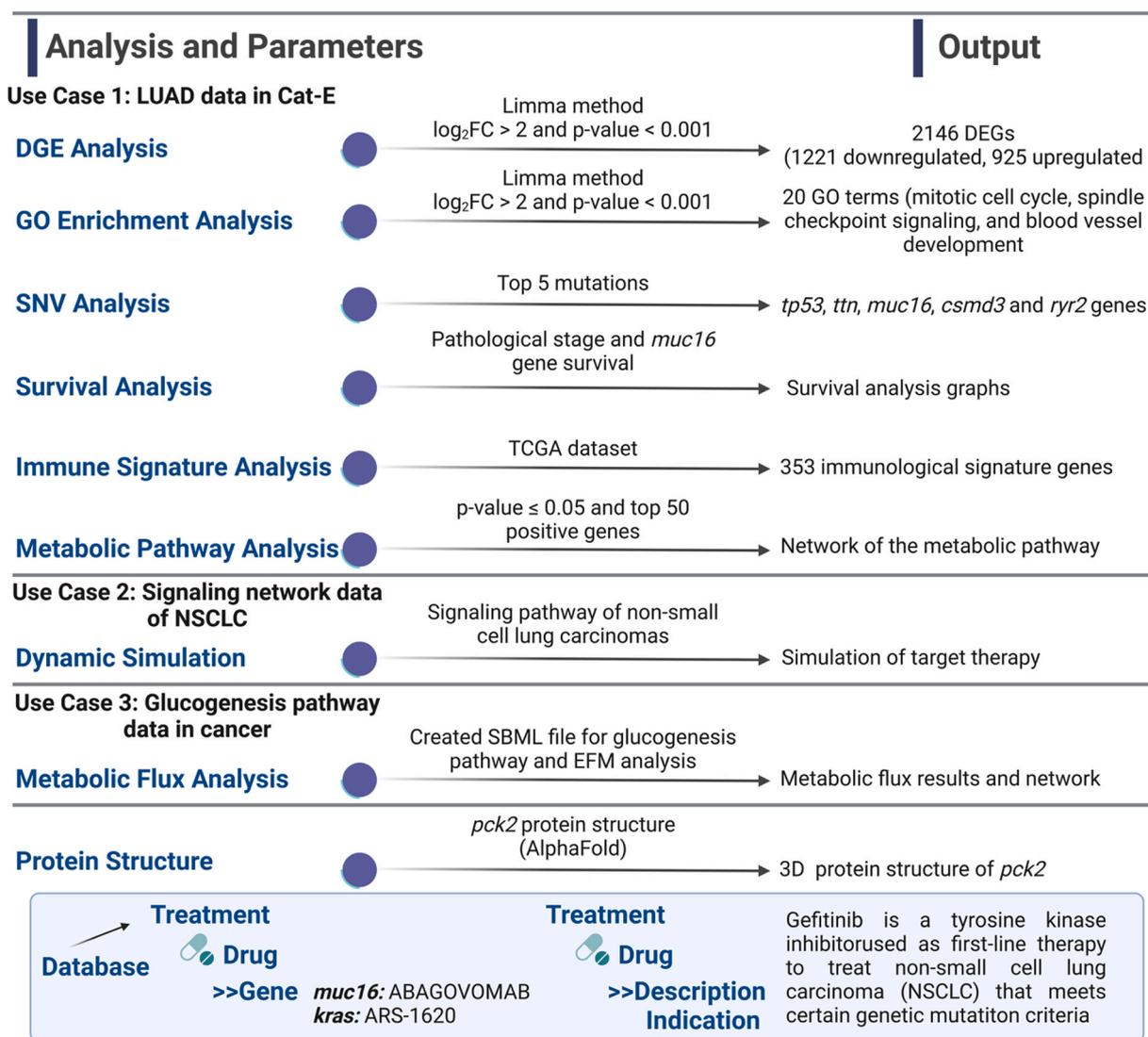
Fig. 2. Application of Cat-E for cancer research. Summary of analytical parameters and outputs for the LUAD case studies performed using Cat-E. Potential targets for anti-cancer drug action are explored using the databases in Cat-E.

illustrating the top 20 GO terms pertinent to LUAD after filtering, covering biological processes, molecular functions, and cellular components ($\log_2$FC > 2 and p-value < 0.001; Fig. 3C).

3.1.3. Analysis of single nucleotide variations (SNVs)

Cat-E allows detailed analysis of SNVs, which are significant genetic alterations affecting cellular function and disease progression. The analysis of somatic mutation patterns in 616 patients with LUAD by Cat-E (Fig. 3D) elucidates the genes *tp53, ttn,* and *muc16* with the highest mutation frequencies of 50%, 43%, and 41%, respectively (Supplementary Fig. 1). *Muc16* can be a significant mutation hotspot, as depicted in the lollipop plot (Fig. 3D). Missense mutations are the most common variant classification in LUAD patients. The most prevalent mutations in the variant category of the single nucleotide are cytosine to adenine transversions.

3.1.4. Survival and immune signature analyses

Cat-E provides advanced survival analysis tools for visualizing survival outcomes using clinical data stratified by tumour classifications. Analysis of survival data from LUAD patients, stratified by pathologic staging, unveils distinct survival trends. Notably, the stage 1B cohort exhibits the longest survival duration compared to the other trajectories (as depicted in Fig. 4A). Furthermore, analysis of expression levels of

muc16, a frequently regulated and mutated gene, indicates no significant correlation with survival outcomes (Fig. 4B).

Moreover, Cat-E facilitates the study of immunological signatures by arranging crucial immune genes in networked and tabular layouts. We conducted a correlation network analysis to investigate gene interactions within immunological clusters specific to LUAD. Cat-E systematically organized and illustrated basic immunological signatures (ImSig) and their relationships with various types of immune cells.

3.1.5. Metabolic pathway analysis result

The Cat-E framework incorporates the functionality for metabolic pathway analysis, enabling the identification of canonical metabolic pathways, with particular emphasis on those exhibiting an adjusted p-value of ≤ 0.05. Additionally, it allows users to examine specific reactions along with their corresponding enzymes. As an illustrative example, we constructed a metabolic pathway network within Cat-E using LUAD data, focusing on the top 50 positively associated genes (see Fig. 4C). Cat-E enabled the identification of active metabolic pathways in the LUAD data, as depicted in Fig. 4C where increased lipid metabolism is evident.
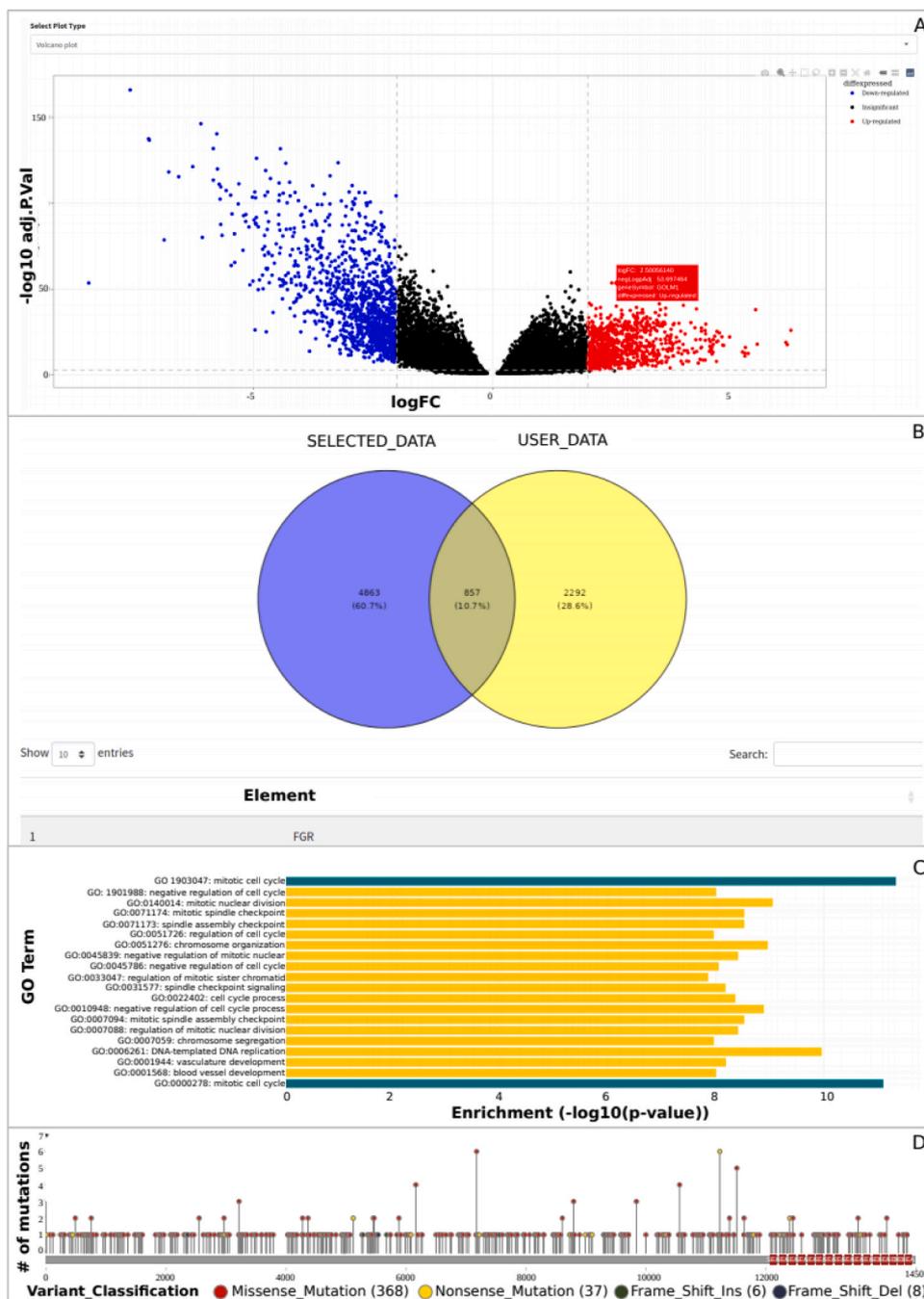
Fig. 3. A. Volcano plot illustrating DGEs in LUAD (the black, blue, and red circles represent nonsignificant, down-regulated, and up-regulated genes, respectively); **B.** A Venn diagram showing the comparison between the LUAD data within Cat-E (Selected_Data) and the DGE LUAD data in the GEPIA2 (User_Data); **C.** A bar graph displaying the top 20 GO terms obtained from enrichment analysis; **D.** Visual representation of mutation positions within the *muc16* gene in LUAD.

3.2. Use case 2: Dynamic simulation of gene regulatory network analysis

In the second use case, we employed the non-small cell lung carcinomas (NSCLC) dataset from the study investigating the role of epithelial-to-mesenchymal transition (EMT) in various conditions [13]. Subsequently, dynamic simulations of gene regulatory networks were created using the integrated Jimena module [47]. Based on dynamic simulation outcomes using the provided signaling network [13], a combination of targeted drugs involving ARS-1620 and alisertib demonstrated efficacy in overcoming resistance in NSCLC with KRAS mutation, particularly in cell lines characterized by high c-MYC expression levels. The XML and TXT formats of this model are uploaded to our repository in GitHub (https://github.com/salihoglu/Cat-E) to

facilitate reproduction of this analysis.

To enhance the usability of the application, we have developed a dedicated database. Moreover, with Cat-E, users can generate the required files for the Jimena module [47]. This capability enables dynamic simulation of pathway data established in previous research. Our integrated system streamlines the creation of corresponding TXT files and expedites the extraction of relevant gene data from the database.

3.3. Use case 3: Metabolic flux analysis: gluconeogenesis pathway in cancer

Cat-E provides functionality for generating metabolic SBML files for users lacking pre-existing ones, using the integration between COBRA
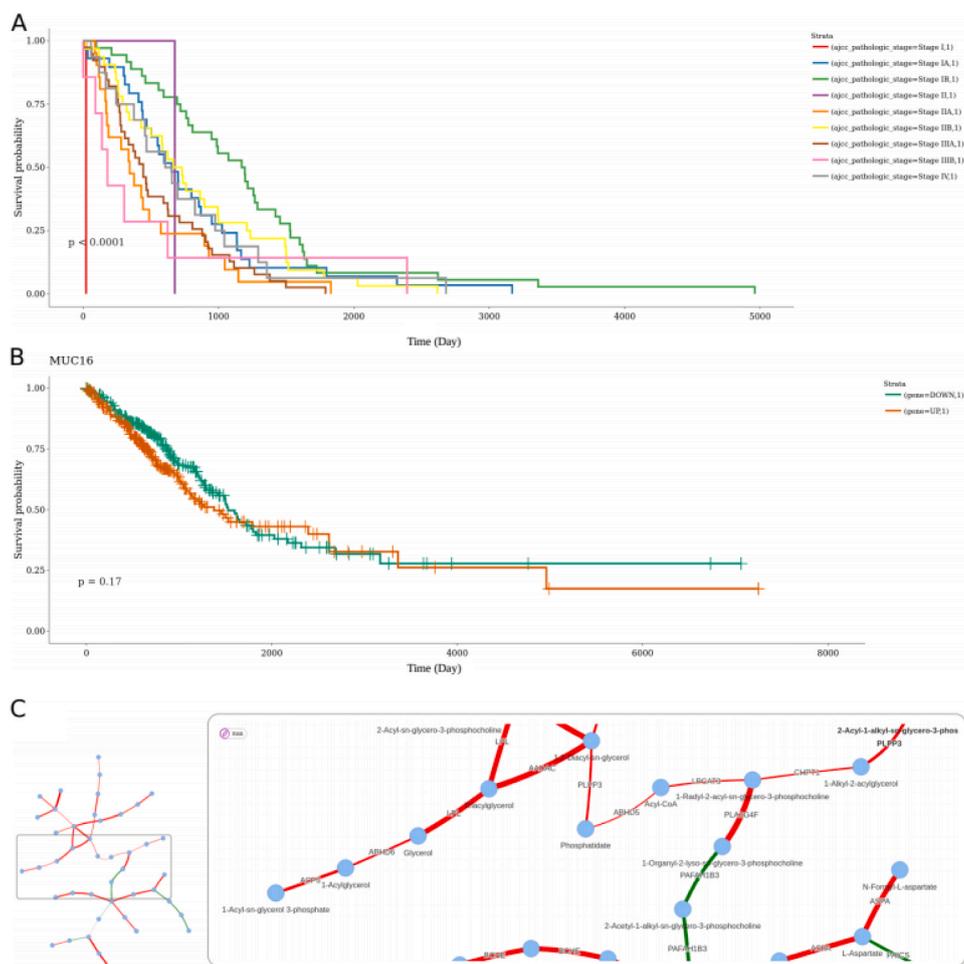
Fig. 4. A. Survival analysis based on pathological staging in LUAD clinical data, with each color representing a distinct pathological stage (Red: Stage1, Blue: Stage1A, Green: Stage1B, Purple: Stage2, Orange: Stage2A, Yellow: Stage2B, Brown: Stage3A, Pink: Stage3B, Grey: Stage4); **B.** Comprehensive analysis of overall survival regarding the expression levels (Green: down-regulation, Orange: up-regulation) of *muc16* in LUAD tumor samples; **C.** Identification of active metabolic pathways within LUAD data, illustrated by a red line indicating $log_2FC> 0$ and a green line indicating $log_2FC< 0$.

and d3flux tools (see Fig. 5A-B). Fig. 5B illustrates the exploration of the gluconeogenesis pathway in cancer cells [53], initiated through an EFM analysis. Notably, our metabolic flux evaluation revealed a significant increase in the flux of *pck2* (phosphoenolpyruvate carboxykinase 2) in LUAD cancer (see Fig. 5C). This finding is particularly noteworthy in light of the concurrent upregulation of the *pck2* gene observed in the DGE dataset. Our investigation not only addresses the intricate dynamics of the gluconeogenesis pathway, but also elucidates the molecular basis of this process, highlighting the potential implications of *pck2* upregulation in LUAD.

3.4. Detailed cancer targets: example of PCK2 protein structure

The integration of AlphaFold [49,50] provides easy and effective structural analysis on biological macromolecules, which are crucial for understanding potential therapeutic targets. As an example, we focus on the *pck2* (Q16822) protein. Fig. 6 shows an interactive model of *pck2* for protein structure prediction, generated using AlphaFold. The specific amino acids can be selected, as demonstrated in yellow (labeled as 'selection: 1–20'; Fig. 6), by the illustration of the first 20 amino acids of the protein sequence. Furthermore, the interface provides options for adjusting colors and offers features such as displaying ligands, rotating the model, and formatting options, thereby providing valuable insights into the structural characteristics.

Through these functions within Cat-E, users can explore the predictive capabilities of AlphaFold by providing a UniProt ID and can adjust

the structural characteristics.

Numerous opportunities for further exploration in cancer research are possible, based on the availability of suitable omics data. These include detailed analyses of cancer pathways and investigations into combination therapies, as evidenced by our previous studies on combinatorial-targeted treatment in the NSCLC with aggressive KRAS-Biomarker signatures [13], modeling of stem cell differentiation [54], and further usage of Cat-E tools for stratifying cancer types in colorectal cancer [55] or lung cancer [56].

3.5. Obtaining drug combination and targeting strategies insights using Cat-E

The comprehensive analyses conducted using Cat-E provide insights into the molecular landscape of cancer, offering researchers a wealth of data to explore potential therapeutic targets. The analysis shows that the upregulated *muc16* gene is correlated with abagovomab, gefitinib, and bevacizumab drug treatments, as indicated in the gene-drug database integrated within Cat-E (Table 1). These data can be retrieved from two different sources, directly from the Drug_>genes (e.g.muc16) or via the Drug-> Description ->Indication (e.g. lung cancer).

Moreover, metabolic flux analysis revealed significant alterations in the gluconeogenesis pathway, particularly concerning the upregulation of the Pycr family. These findings emphasize the potential usage of targeting metabolic pathways in lung cancer to find new routes of cancer cell growth prevention. Furthermore, structural analysis of proteins
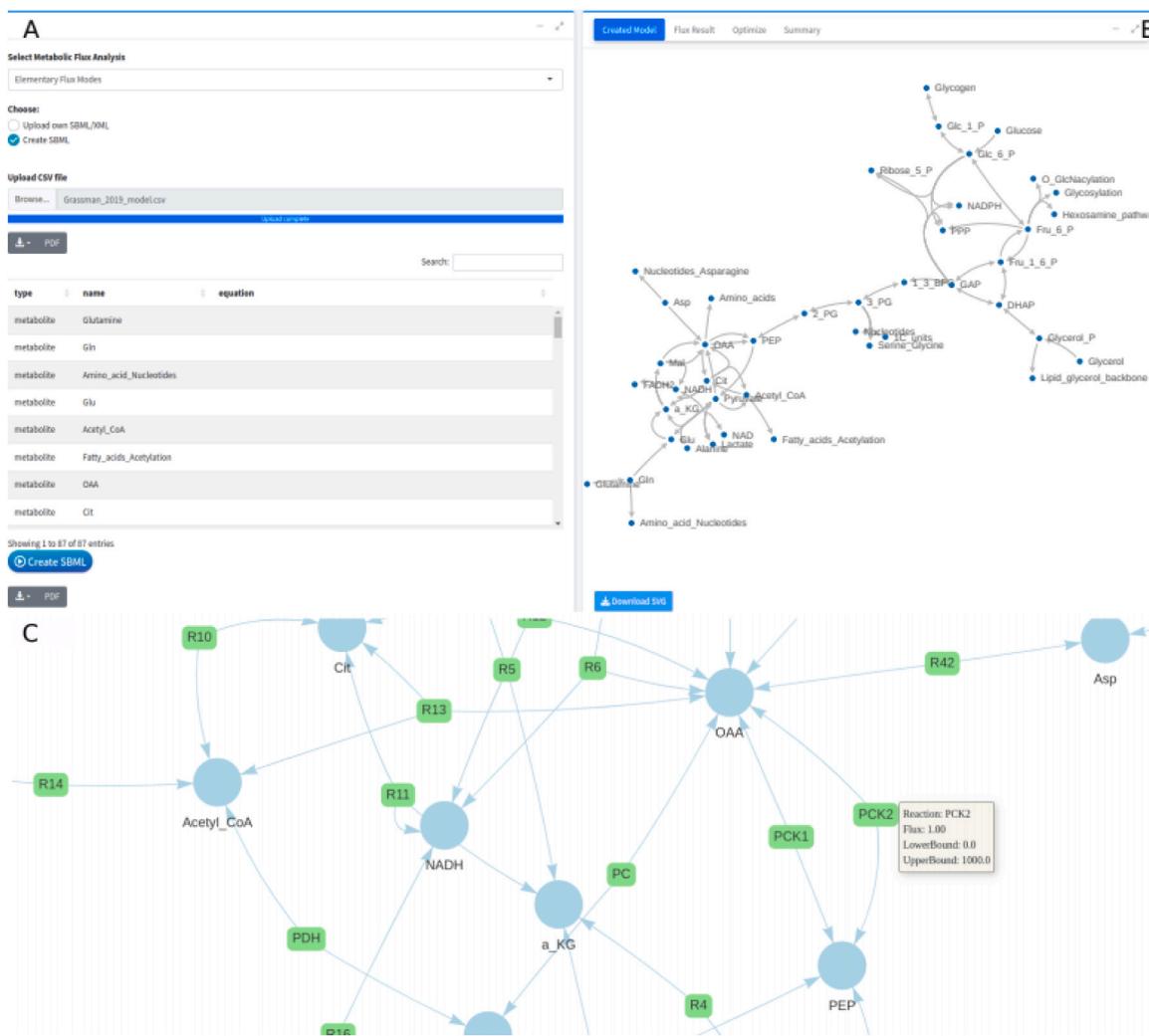
Fig. 5. Analysis of the metabolic flux in LUAD cancer cells. **A.** Module for loading or generating the necessary.sbml file essential for conducting metabolic flux analysis; **B.** Representation of the generated metabolic model (additional tabs furnish results of flux analysis, optimized outcomes, and summary data); **C.** Computed flux values for the reactions generated, presented as a network.

using AlphaFold offers insights into potential druggable targets, providing a foundation for the design efforts of structure-based drugs.

To further enhance cancer protein and drug target evaluation, especially regarding oncogenes, tumor drivers, and failure of different apoptosis pathways (e.g. mutation or loss of p53), researchers can use Cat-E to create dynamic signaling networks based on the obtained data. By simulating these networks and testing relevant drugs within the dynamic context, researchers can refine their understanding of signaling pathways implicated in cancer progression and identify promising candidates for therapeutic intervention. Cat-E offers a systematic dynamic framework for exploring and validating potential cancer targets, as a critical step in advancing precision medicine, by looking at individual tumor drivers, driver combinations, signature constellations and even evaluating the effect of specific mutations in the light of omics data to predict successful drug targeting or resistance.

4. Discussion

The Cat-E web tool was explicitly developed to examine proteins and protein cascades involved in cancer as promising drug targets against a background of alternative strategies (oncolytic viruses, immune strategies, cytostatics) and provides user-friendly access to crucial databases. The user can upload their data in various formats, including DGE data, as well as other omics and functional data to help interpret differences in

DGEs. Cat-E offers a wide variety of analytical tools for cancer targeting-related studies, capable of conducting nine different types of analyses to meet the diverse requirements of the cancer research community, such as biologists, oncologists, and clinicians.

Cat-E can analyze TCGA-GTEx data for 33 different cancer types and perform DGE analysis on GEO data using the user's GSE ID, as well as presentation of immune gene signatures. Cat-E offers datasets available through GEO via a user-friendly web tool tailored for individuals lacking advanced programming skills. This web application provides a thorough and standardized protocol of gene expression analysis for effectively accessing GEO datasets. The analysis results can be stored in a CSV file format. It simplifies online metabolic flux analysis by smoothly incorporating the CNApy tool. Moreover, Cat-E is a crucial tool for categorizing research on various cancer drug targets and targeting strategies, such as cytostatic drugs, CAR-T cells, cancer-targeting bispecific antibodies, and oncolytic virus methods, in light of all available omics data including clinical trial data, as available (e.g. lists of long- and short-survival-related genes). This categorization is essential for evaluation of cancer targeting strategies, as Cat-E assists in navigating not only direct protein structures but also protein and pathway combinations as well as the extensive range of immunomodulatory approaches.

Case studies demonstrate that Cat-E enables rapid assessment of individual genes, their expression, and their significance in cancer. For example, one of the primary genes that is increased in expression in
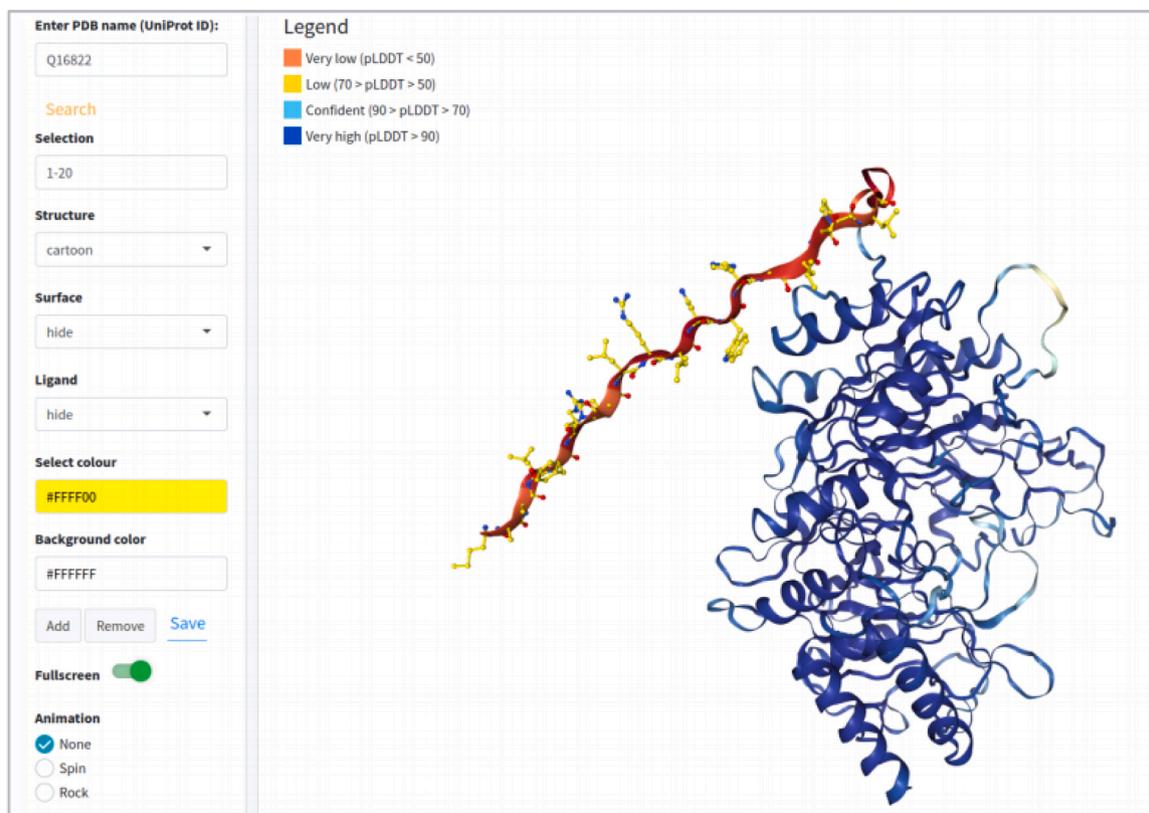
Fig. 6. Interactive model of the *pck2* (Q16822) protein in Cat-E, generated using AlphaFold for protein structure prediction. The visualization highlights the first 20 amino acids of the protein sequence (selection: 1–20), depicted in yellow. The color scheme applied to the protein structure is based on AlphaFold confidence scores, with dark blue representing very high confidence (pLDDT > 90), light blue indicating high confidence (90 > pLDDT > 70), yellow reflecting low confidence (70 > pLDDT > 50), and orange denoting very low confidence (pLDDT < 50). This confidence-based coloring scheme facilitates an in-depth assessment of structural predictions.

Table 1
The list of the drugs associated with *muc16* gene in the treatment of lung cancer.

Drug name	Interaction claim source	Drug concept ID
ABAGOVOMAB	TdgClinicalTrial, TTD ChemblInteractions	CHEMBL1742981
GEFITINIB	Drugbank online	CHEMBL939
BEVACIZUMAB	Drugbank online	CHEMBL1201583
OREGOVOMAB	TdgClinicalTrial ChemblInteractions	CHEMBL2107917
TAMOXIFEN	NCI	CHEMBL83
DOCETAXEL	NCI	CHEMBL92
CYCLOSPORINE	NCI	CHEMBL160
TOPOTECAN	NCI	CHEMBL84
ETOPOSIDE	NCI	CHEMBL44657
DMUC-5754A (SOFITUZUMAB VEDOTIN)	TTD	CHEMBL3545372

NCI: National Cancer Institute, TTD: Therapeutic Target Database

LUAD is *muc16*. This gene was identified as a significant hotspot for mutations through SNV analysis using Cat-E. On a note of caution, Cat-E survival plots do not exhibit a straightforward and evident correlation between upregulation and survival. However, the significance of *muc16* in cancer has been documented in the literature [59].

Existing tools such as GEPIA2 [3], GENAVi [5], GNOSIS [57], HPVTIMER [58], and the ShinyGATOM app [12] enable expression, survival, and correlation analyses using TCGA and GTEx data, as well as metabolic pathway analysis. Cat-E distinguishes itself by providing a wider variety of analysis tools and databases all in one place. Aside from analyzing cancer omics data, multiple network analysis tools are accessible for investigating cancer and evaluating protein-based cancer

drug targets for comparison to other anti-cancer strategies (immune cells, immune modulation, oncolytic virus). These tools encompass general, regulatory, and metabolic network analyses and are equipped to examine user-provided data sets in the required format. Cat-E explores targets and target pathways regarding oncolytic viruses, targeting cancer proteins and drug evaluation, combined effects of kinase inhibitors and tumor driver inhibitors, cytostatic therapy, immunomodulatory therapies like CAR-T cells, as well as interactive protein structure examination and drug targeting. Cat-E does not function as a clinical study nurse tool nor is it intended for use in clinic or for monitoring or assessing patient therapy directly; instead, it concentrates on molecular interactions and mechanisms.

The Cat-E platform will enhance its robust capabilities in the future by incorporating additional omics network analysis tools, embracing open-source development and integrating with advanced computational models such as large language models (LLMs). Cat-E is positioned to narrow the gap between bench research and clinical research by broad evaluation of targeting strategies and protein targets.

5. Conclusions

The Cat-E (**Ca**ncer **T**arget **E**xplorer) web tool is a sophisticated platform that provides a multidimensional view of omics data to evaluate various cancer protein targeting strategies as well as immunological and oncolytic intervention strategies in cancer research. It includes differential gene expression, metabolic flux, single nucleotide variations, visualization of predicted and experimental protein structure, drug interactions, and exploration. As shown for lung adenocarcinoma (LUAD) datasets, Cat-E demonstrates its utility in enabling users to explore genetic variations, identify diagnostic markers, and propose

therapeutic targets across various cancer types. This tool is especially valuable for experimental biologists, clinical scientists, and students, bridging the gap between complex bioinformatics and extensive target and pathway evaluation. By simplifying access to as well as analysis of complex data sets, Cat-E enhances our understanding of cancer biology, including the evaluation of personalized cancer protein targeting strategies including combinations, immune modulation and oncolysis.

No generative AI was used in writing of this manuscript

The authors state clearly that they **did not use** any generative artificial intelligence (AI) and AI-assisted technologies in the writing process.

Author contributions

RS programmed, tested, and implemented Cat-E, supervised by EB and TD. CL and JB configured Nginx. GD provided cancer insights and data. All authors edited and polished the manuscript together and approved the final version.

CRediT authorship contribution statement

Thomas Dandekar: Conceptualization, Formal analysis, Investigation, Supervision, Writing – review & editing. **Elena Bencurova:** Conceptualization, Formal analysis, Investigation, Supervision, Writing – review & editing. **Rana Salihoglu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Johannes Balkenhol:** Methodology, Writing – review & editing. **Gudrun Dandekar:** Investigation, Validation, Writing – review & editing. **Chunguang Liang:** Methodology, Writing – review & editing.

Data and code availability

All data, including access information for the software and genetic constructs, are fully available in the manuscript and the Supplementary Information. Specifically, all code and data can be accessed at https://github.com/salihoglu/Cat-E.

Acknowledgements

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.03.024.

References

[1] https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] Sun W, Duan T, Ye P, Chen K, Zhang G, Lai M, et al. Tsvdb: A web-tool for tcga splicing variants analysis. BMC Genom 2018;19:1–7. https://doi.org/10.1186/S12864-018-4775-X/TABLES/1.

[3] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 2019;47(W1):W556–60.

[4] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Sci Signal 2013:6. https://doi.org/10.1126/SCISIGNAL.2004088.

[5] Reyes ALP, Silva TC, Coetzee SG, Plummer JT, Davis BD, Chen S, et al. Genavi: A shiny web application for gene expression normalization, analysis and visualization. BMC Genom 2019;20:1–9. https://doi.org/10.1186/S12864-019-6073-7/TABLES/1.

[6] Ge SX, Jung D, Jung D, Yao R. Shinygo: A graphical gene-set enrichment tool for animals and plants. Bioinformatics 2020;36:2628–9. https://doi.org/10.1093/BIOINFORMATICS/BTZ931.

[7] Salihoglu R, Srivastava M, Liang C, Schilling K, Szalay A, Bencurova E, et al. Prosimat: Protein network simulation and design tool. Comput Struct Biotechnol J 2023;21:2767–79. https://doi.org/10.1016/J.CSBJ.2023.04.023.

[8] Hunt GP, Grassi L, Henkin R, Smeraldi F, Spargo TP, Kabiljo R, Koks S, Ibrahim Z, Dobson RJ, Al-Chalabi A, Barnes MR. GEOexplorer: a webserver for gene expression analysis and visualization. Nucleic Acids Res 2022;50(W1):W367–74.

[9] Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. Nucleic Acids Res 2015;43(D1):D968–73.

[10] Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. Intogen-mutations identifies cancer drivers across tumor types. Nat Methods 2013 10:11 2013;10:1081–2. https://doi.org/10.1038/nmeth.2642.

[11] Noecker C, Eng A, Muller E, Borenstein E. Mimosa2: a metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data. Bioinformatics 2022;38:1615–23. https://doi.org/10.1093/BIOINFORMATICS/BTAC003.

[12] Emelianova, M., Gainullina, A., Poperechnyi, N., Loboda, A., Artyomov, M., Sergushichev, A. Shiny gatom: omics-based identification of regulated metabolic modules in atom transition networks. academicoupcom 2022;50:W690–W696. doi: 10.1093/nar/ gkac427.

[13] Peindl M, Gottlich C, Crouch S, Hoff N, Luttgens T, Schmitt F, et al. Emt, stemness, and drug resistance in biological context: A 3d tumor tissue/in silico platform for analysis of combinatorial treatment in nsclc with aggressive kras-biomarker signatures. Cancers 2022:14. https://doi.org/10.3390/CANCERS14092176.

[14] Lathwal A, Kumar R, Raghava GP. Ovirustdb: A database of oncolytic viruses for the advancement of therapeutics in cancer. Virology 2020;548:109–16. https://doi.org/10.1016/J.VIROL.2020.05.016.

[15] Cook HV, Doncheva NT, Szklarczyk D, Mering CV, Jensen LJ. Viruses. string: a virus-host protein-protein interaction database. mdpicom 2018:10. https://doi.org/10.3390/v10100519.

[16] Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The intact molecular interaction database in 2012. D841–D846 Nucleic Acids Res 2012;40. https://doi.org/10.1093/NAR/ GKR1088.

[17] Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, Zhang F. The BioGRID interaction database: 2019 update. Nucleic Acids Res 2019;47(D1):D529–41.

[18] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M. TCGAbiolinks: an R/ Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016;44(8):e71.

[19] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, Chen J, Jiang W, Yang K, Ou Q, Li X. CellMarker 2.0: an updated database of manually curated cell markers in human/ mouse and web tools based on scRNA-seq data. Nucleic Acids Res 2023;51(D1):D870–6.

[20] Osorio D, Kuijjer ML, Cai JJ. rPanglaoDB: an R package to download and merge labeled single-cell RNA-seq data from the PanglaoDB database. Bioinformatics 2022;38(2):580–2.

[21] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46(D1):D1074–82.

[22] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47(D1):D1102–9.

[23] Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, Krysiak K, Pan D, McMichael JF, Eldred JM, Walker JR. DGIdb 2.0: mining clinically relevant drug–gene interactions. Nucleic Acids Res 2016;44(D1):D1036–44.

[24] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 2007;36(suppl_1). D684-8.

[25] Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The clinicaltrials.gov results database — update and key issues. N Engl J Med 2011;364:852–60. https://doi.org/10.1056/NEJMSA1012065.

[26] Zhang Y, Yao Y, Chen P, Liu Y, Zhang H, Liu H, et al. Checkpoint therapeutic target database (ckttd): The first comprehensive database for checkpoint targets and their modulators in cancer immunotherapy. J Immunother Cancer 2020:8. https://doi.org/10.1136/JITC-2020-001247.

[27] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for rna-sequencing and microarray studies. e47–e47 Nucleic Acids Res 2015;43. https://doi.org/10.1093/NAR/GKV007.

[28] Robinson, M., McCarthy, D., bioinformatics, G.S., undefined 2010, edger: a bioconductor package for differential expression analysis of digital gene expression

data. academicoupcomMD Robinson, DJ McCarthy, GK Smythbioinformatics, 2010●academicoupcom 2010;26:139–140. doi:10.1093/bioinformatics/btp616.

[29] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/ bioconductor package biomart. Nat Protoc 2009;4: 1184–91. https://doi.org/10.1038/NPROT.2009.97.

[30] Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 2007;23(14):1846–7.

[31] Thiele S, Kamp AV, Bekiaris PS, Schneider P, Klamt S. Cnapy: a cellnetanalyzer gui in python for analyzing and designing metabolic networks. Bioinformatics 2022; 38:1467–9. https://doi.org/10.1093/BIOINFORMATICS/BTAB828.

[32] Loboda, A., Poperechnyi, N., Alvarez-Miranda, E., Sinnl,M., Sergushichev, A. mwcsr: Solvers for Maximum Weight Connected Subgraph Problem and Its Variants; 2023. R package version 0.1.7.

[33] Ushey, K., Allaire, J., Tang, Y. reticulate: Interface to 'Python'; 2023. Https:// rstudio.github.io/reticulate/,https://github.com/rstudio/reticulate.

[34] Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: constraints-based reconstruction and analysis for python. BMC Syst Biol 2013;7:1–6.

[35] Alexa A, Rahnenführer J. Gene set enrichment analysis with topGO. Bioconductor Improv 2009;27:1–26.

[36] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. Innovation 2021;2(100141). https:// doi.org/10.1016/j.xinn.2021.100141.

[37] Luo Weijun, Brouwer Cory. Pathview: an r/bioconductor package for pathway-based data integration and visualization. Bioinformatics 2013;29(14):1830–1. https://doi.org/10.1093/bioinformatics/btt285.

[38] Therneau, T.M. A Package for Survival Analysis in R; 2023. URL: https://CRAN.R-project.org/package=survival; r package version 3.5–7.

[39] Kassambara, A., Kosinski, M., Biecek, P. survminer: Drawing Survival Curves using 'ggplot2'; 2021. R package version 0.4.9.

[40] Almende B.V. and Contributors, Thieurmel, B. visNetwork: Network Visualization using 'vis.js' Library; 2022. R package version 2.1.2.

[41] Li R, Bilal U. Interactive web-based data visualization with r, plotly, and shiny (carson sievert). Biometrics 2021;77:776–7. https://doi.org/10.1111/BIOM.13474.

[42] Wilkinson L. ggplot2: Elegant graphics for data analysis by wickham, h. Biometrics 2011;67:678–9. https://doi.org/10.1111/J.1541-0420.2011.01616.X.

[43] Nirmal AJ, Regan T, Shih BB, Hume DA, Sims H, Freeman TC. Immune cell gene signatures for profiling the microenvironment of solid tumors. AACR 2018;6: 1388–400. https://doi.org/10.1158/2326-6066.CIR-18-0342.

[44] Fukushima A. DiffCorr: an R package to analyze and visualize differential correlations in biological networks. Gene 2013;518(1):209–14.

[45] Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018;28: 1747–56. https://doi.org/10.1101/GR.239244.118.

[46] Guo X, Zhang B, Zeng W, Zhao S, Ge D. G3viz: an R package to interactively visualize genetic mutation data using a lollipop-diagram. Bioinformatics 2020;36 (3):928–9.

[47] Karl S, Dandekar T. Jimena: Efficient computing and system state identification for genetic regulatory networks. BMC Bioinforma 2013:14. https://doi.org/10.1186/1471-2105-14-306.

[48] Valdeolivas, A., Turei, D., Gabor, A. Omnipathr: client for the omnipath web service. 2019. Bioconductor Package.

[49] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. Nature 2021;596:583–9. https://doi.org/10.1038/S41586-021-03819-2.

[50] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. academicoupcomM Varadi. academicoupcom 2021;50:D439–D444Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, editors. O Str Acids Res 2022. https://doi.org/10.1093/nar/gkab1061.

[51] van der Velden, N. Interactive 3d visualization of molecular structures r package nglviewer version 1.3.1. 2021.

[52] Rego N, Koes D. 3dmol.js: molecular visualization with webgl. Bioinformatics 2015;31:1322–4. https://doi.org/10.1093/BIOINFORMATICS/ BTU829.

[53] Grasmann G, Smolle E, Olschewski H, Leithner K. Gluconeogenesis in cancer cells – repurposing of a starvation-induced metabolic pathway? Biochim Et Biophys Acta - Rev Cancer 2019;1872:24–36. https://doi.org/10.1016/J.BBCAN.2019.05.006.

[54] Kaltdorf M, Breitenbach T, Karl S, Fuchs M, Kessie DK, Psota E, et al. Software jimenae allows efficient dynamic simulations of boolean networks, centrality and system state analysis. Sci Rep 2023 13 2023;1(13):1–16. https://doi.org/10.1038/s41598-022-27098-7.

[55] Baur F, Nietzer SL, Kunz M, Saal F, Jeromin J, Matschos S, Linnebacher M, Walles H, Dandekar T, Dandekar G. Connecting cancer pathways to tumor engines: A stratification tool for colorectal cancer combining human in vitro tissue models with boolean in silico models. Cancers 2019;12(1):28.

[56] Gottlich C, Kunz M, Zapp C, Nietzer SL, Walles H, Dandekar T, et al. A combined tissue-engineered/in silico signature tool patient stratification in lung cancer. Mol Oncol 2018;12:1264–85. https://doi.org/10.1002/1878-0261.12323.

[57] King L, Flaus A, Coughlan S, Holian E, Golden A. Gnosis: an r shiny app supporting cancer genomics survival analysis with cbioportal. HRB Open Res 2022:5. https://doi.org/10.12688/HRBOPENRES.13476.2.

[58] Liu L, Xie Y, Yang H, Lin A, Dong M, Wang H, et al. Hpvtimer: A shiny web application for tumor immune estimation in human papillomavirus-associated cancers. iMeta 2023;2:e130. https://doi.org/10.1002/IMT2.130.

[59] Lee DH, Choi S, Park Y, Jin HS. Mucin1 and Mucin16: therapeutic targets for cancer therapy. Pharmaceuticals 2021;14(10):1053.