# Artificial Intelligence uncovers carcinogenic human metabolites

Aayushi Mittal[1,†], Vishakha Gautam[1,†], Roshan S[2#], Sakshi Arora[1#], Sheetanshu Saproo[7#], Ria Gupta[1], Sanjay Kumar Mohanty[1], Prakriti Garg[1], Anmol Aggarwal[1], Padmasini R[1], Nilesh Kumar Dixit[1], Vijay Pal Singh[6], Anurag Mehta[5], Juhi Tayal[5], Srivatsava Naidu[7], Debarka Sengupta[1,2,3,4,*], and Gaurav Ahuja[1,*]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India

[2]Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.

[3]Centre for Artificial Intelligence, Indraprastha Institute of Information Technology, New Delhi, India

[4]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia

[5]Rajiv Gandhi Cancer Institute & Research Centre, Sir Chotu Ram Marg, Rohini Institutional Area, Sector 5, Rohini, New Delhi, Delhi 110085, India

[6]CSIR-Institute of Genomics & Integrative Biology, South Campus, Mathura Road, Opp: Sukhdev Vihar Bus Depot, New Delhi 110025, India

[7]Department of Bio-Medical Engineering, Indian Institute of Technology Ropar, Rupnagar, Punjab - 140001, India

[†]co-first authors

[#]shared second authors

Correspondence:

Gaurav Ahuja; Email: gaurav.ahuja@iiitd.ac.in

Debarka Sengupta; Email: debarka@iiitd.ac.in

## ABSTRACT

The genome of a eukaryotic cell is often vulnerable to both intrinsic and extrinsic threats due to its constant exposure to a myriad of heterogeneous chemical compounds. Despite the availability of innate DNA damage repair pathways, some genomic lesions trigger cells for malignant transformation. Accurate prediction of carcinogens is an ever-challenging task due to the limited information about *bonafide* (non)carcinogens. This, in turn, constrains the generalisability of such models. We developed a novel ensemble classifier (Metabokiller) that accurately recognizes carcinogens by quantitatively assessing their chemical composition as well as potential to induce proliferation, oxidative stress, genotoxicity, alterations in epigenetic signatures, and activation of anti-apoptotic pathways, therefore obviates the need for *bonafide* (non)carcinogens for training model. Concomitant with the carcinogenicity prediction, it also reveals the contribution of the aforementioned biochemical processes in carcinogenicity, thereby making the proposed approach highly interpretable. Metabokiller outwits existing best practice methods for the carcinogenicity prediction task. We used Metabokiller to decode the cellular endogenous metabolic threats by screening a large pool of human metabolites and identified putative metabolites that could potentially trigger malignancy in normal cells. To cross-validate our predictions, we performed an array of functional assays and genome-wide transcriptome analysis on two Metabokiller-flagged, and previously uncharacterized human metabolites by using *Saccharomyces cerevisiae* as a model organism and observed larger synergy with the prediction probabilities. Finally, the carcinogenicity potential of these metabolites was evaluated using a malignancy transformation assay on human cells.

**Keywords:** Yeast, human, carcinogens, DNA damage, ROS, Deep Learning, Artificial intelligence

## INTRODUCTION

Eukaryotic cells encounter a large number of structurally, and functionally distinct chemicals in their life cycle (*1*–*3*). Such interactions often trigger dysregulation of cellular processes, resulting in loss of cellular homeostasis (*1*, *2*, *4*). Some of these chemical moieties alter the critical components of a cell, such as its genome or the surveillance mechanisms that ensure its integrity, leading the cells to malignant transformation (*4*–*7*). Such a class of chemicals is called carcinogens. Moreover, due to the ever-increasing rate of cancer incidence worldwide, carcinogens are emerging as a major health hazard (*8*). Notably, only 5-10% of the cancer types are heritable in nature, whereas, the majority of the cancers are caused due to exposure to carcinogens (*9*). Mechanistic insights about their mode of action are still under exploration, however, multiple studies revealed that carcinogens also impact the genetic material (DNA) of a cell and globally alter the key cellular machinery such as its epigenome, transcriptome, proteome, and metabolome (*4*, *10*, *11*). Recent reports support the notion that a chemical carcinogen could induce carcinogenicity by directly impairing the epigenome, interfering with the DNA damage repair response pathway, or activating the anti-apoptotic pathways (*4*, *10*, *11*). Moreover, it has also been observed that carcinogens trigger an alteration in the cellular microenvironment that leads to energy metabolic dysfunction and induces carcinogenesis (*12*–*14*). All these indicate that the underlying mechanisms of the carcinogens' mode-of-action are rather convoluted. As such, identification and characterization of carcinogens are imperative to prevent human exposure to these threats. Present approaches to zero in on carcinogens are contingent on expensive (up to $4 million per compound) and time-consuming (more than two years) animal model-based approaches (*15*, *16*). Strikingly, an average validation involves harvesting of ~800 animals, inflicting ethical concerns (*15*, *16*). While *in vivo* experiments offer indisputable identification of carcinogens, a helping hand of Artificial Intelligence could accelerate the pre-screening of the ever-expanding space of chemicals that includes new drugs, chemicals, and industrial by-products. To date, numerous computational methods have been proposed for carcinogenicity prediction. In most cases, quantitative structure-activity relationship (QSAR) models have been used for the classification of carcinogens and non-carcinogens (*17*–*21*). Moreover, structural

alert-based expert systems are also used where the carcinogenic potential of a compound is estimated using two-dimensional structural similarities (*22*, *23*). QSAR models offered decent predictive performance due to their ability to detect specific functional groups that are reported or linked to carcinogenicity. Some of the prominent functional groups include nitro chemicals, aromatic amines, polycyclic aromatic hydrocarbons, and polychlorinated biphenyls (*18*, *19*, *24*, *25*). To date, all predictive models for carcinogenicity prediction are supervisory in nature and use a limited number of known carcinogens and non-carcinogens for model training. Given the complexity of the chemical space, this approach fails to generalize unseen chemicals with discordant structural properties. With the advancement in functional assays, it has been established that a potential carcinogen might induce cellular proliferation (*26*, *27*), genotoxicity, oxidative stress response, anti-apoptotic response (*27*), and epigenetic alterations (*28*). Additionally, a large number of carcinogens are known to be electrophilic in nature (*29*). By leveraging these biochemical properties, we developed Metabokiller, a method that utilizes the biochemical properties-based ensemble model for carcinogenicity prediction, and therefore, obviates the need for known carcinogens or non-carcinogens in model training. By using Metabokiller we delineate a list of human metabolites that possess carcinogenic properties. We selected two previously uncharacterized human metabolites that were predicted as potential carcinogens using Metabokiller and experimentally validated its predictions using an array of functional assays and deep RNA sequencing. In summary, the present work demonstrates a novel computational approach leveraging carcinogen-specific biochemical properties to accurately predict carcinogenicity.

## RESULTS

### Carcinogen-independent Ensemble Model for Carcinogenicity prediction

Cancer cells are highly proliferative, and possess altered epigenetic signatures, elevated reactive oxygen species (ROS) levels, and activated anti-apoptotic pathways (*30*). Moreover, it has been observed that most carcinogens, if not all, are known to be electrophilic in nature and hence are genotoxic (*29*). Unlike other QSAR-based models that largely use information about the known carcinogens and non-carcinogens, we selected a set of well-known biochemical properties to build an ensemble model that evaluates the query chemical for these properties and hence predicts its carcinogenicity potential **(Figure 1A)**. We first manually curated and compiled datasets containing information about chemicals that are reported to positively and negatively/neutrally impair cellular proliferation, genotoxicity, oxidative stress, epigenetic alterations, and anti-apoptotic response. We also compiled a dataset comprising information about *bonafide* electrophiles and non-electrophiles **(Figure 1B, SI Table 1)**. In total, all the six independent datasets contain a total of 5996 and 9613 chemicals, segregated into pro (Class 1) or anti/no (Class 0)-activity, respectively. We next investigated the functional or chemical heterogeneity in each of the biochemical datasets (electrophiles, epigenetic modifications, genotoxicity, oxidative stress, pro-proliferation, and anti-apoptotic properties) by the Principal Component Analysis (PCA) using the Atom Pair Fingerprints as features. We observed a higher heterogeneity in almost all the cases **(Figure 1C)**. Functional group comparisons between the pro (Class 1) or anti-or-no activity (Class 0) groups revealed the selective enrichment or de-enrichment of the twelve tested functional groups in each of these datasets. Collectively, these results indicate the higher degree of functional, compositional, and chemical heterogeneity between the pro or anti/no-activity chemicals for each of these properties **(Figure 1C)**. To build an ensemble model, we next generated six independent classification models featuring the aforementioned biochemical properties. Briefly, for the genotoxicity classification model, we first compiled the chemicals that are classified as mutagens (class 1) or non-mutagens (class 0) in the AMES test **(Figure 1B)**. Next, we built the classification model using the graph convolutional model, a graph-based method that represents chemical structures as graphs and therefore exhibits excellent performances in the

classification tasks. Similar steps of data compilations, followed by the model building were performed for other models such as epigenetic modification (Graph Convolution Network), oxidative stress (Attentive FP), and electrophilic properties (Attentive FP). Of note, in the case of classification models predicting anti-apoptotic and pro-proliferative properties, we used molecular descriptor-based machine learning approaches using Random Forest and Extra Tree classifiers, respectively **(Figure S1A, B, Figure S2B, C)**. Notably, for identifying the best performing computational technique and its associated hyperparameters, we tried and tested multiple algorithms while enabling grid-search-based hyperparameter tuning. Moreover, for some models, the class imbalance problems were handled by oversampling the minority class and the performance of the individual models was tested on the held-out dataset (validation data) using the 10 fold cross-validation **(Figure S1A)**. Once, we obtained the best performance model for each of the six aforementioned biochemical properties **(Figure 1D-F, Figure S2A)**, we next built an ensemble model by combining these six distinct models. This was achieved by evaluating the performances (accuracies) of each of these six models in classifying experimentally-validated carcinogens and non-carcinogens **(Figure 1G)**. Briefly, we used a known dataset (annotated as C1 dataset) comprising a *bonafide* list of experimentally-validated carcinogens and non-carcinogens (**SI Table 2**), and the accuracies of the individual models were calculated. We next assigned weights to each of the six models based on their predicted accuracies on the C1 dataset, and subsequently, the weighted models were used to build the final ensemble model (Metabokiller) **(Figure 1A, G, Figure S1A)**. Notably, the biochemical property-focused Metabokiller by the virtue of its construction offers interpretability. Further nuances of the approach can be found in the Methods section.

**Metabokiller outperformed existing methods for Carcinogenicity prediction and unfolded endogenous metabolic threats**

To test the performance of Metabokiller, we used 2 external datasets, named C2 and C3, respectively **(Figure 2A, SI Table 3,4)**. These datasets contain chemical information about experimentally-validated carcinogens and non-carcinogens. Expectedly, in both of

these datasets, we observed a larger degree of chemical heterogeneity as shown in the 3D-PCA. Functional group comparisons between the *bonafide* carcinogens and non-carcinogens revealed the selective enrichment of ether functional group (ROR) in carcinogens in both the datasets, whereas, strong de-enrichment was observed for primary (RNH2) and Tertiary amines (R3N) in the C3 dataset **(Figure S3A, B)**. We next tested these ground truth datasets on Metabokiller and observed AUC-ROC of 0.940 and 0.848 on C2 and C3 datasets, respectively **(Figure 2B)**. In addition to this, the other performance parameters, such as precision, recall, sensitivity, specificity, and Cohen's Kappa were found to be within the acceptable ranges **(Figure S3C, D)**. These results suggest that in contrast to all the previous methods, Metabokiller exhibits higher accuracy and other performance parameters for the carcinogen classification (*17*, *18*, *21*, *24*, *25*). Next, we performed a comparative analysis of MetaboKiller with three other widely accepted tools or methods, namely, Protox-II (*31*), versatile variable nearest neighbor (vNN) (*32*), and lazy structure-activity relationships (lazar) (*33*, *34*), by using the C2 dataset as an input **(Figure 2C)**. Of note, unlike Metabokiller, all these three methods are built on the information compiled and inferred from the *bonafide* (non)-carcinogens (Protox-II) and/or (non)-mutagens (lazar and vNN) **(Figure S3H)**. Further to avoid any potential bias in the prediction performance due to overlapping chemicals used in the training dataset for each of these models, we performed a filtering step that selectively removed redundant chemicals **(Figure 2C)**. Interestingly, comparative analysis revealed that Metabokiller outperformed all other methods in almost every performance metric. For instance, Metabokiller returned an accuracy of 82.2%, precision of 0.93, F1 score of 0.84, MCC of 0.65, and Cohen's Kappa of 0.64, but a comparable recall of 0.76 with Protox-II **(Figure 2D, Figure S3E-G)**. These results indicate that the selected chemical and biological properties leveraged by Metabokiller truly capture the chemical and functional properties of carcinogens, thereby, enabling their accurate classification.

We next aimed to identify all the cellular metabolites that harbor carcinogenic properties by leveraging Metabokiller predictions. To achieve this, we performed a large-scale *in-silico* screening of the human metabolome. Methodologically, we subjected all human

metabolites, cataloged in the Human Metabolome Database (HMDB) (*35*) on Metabokiller **(Figure 2E)**. HMDB contains information about the small molecule metabolites found in the human body. In total, we projected 1,13,580 human metabolites and calculated their carcinogenicity probabilities. Moreover, as an output, we also obtained the individual probabilities of each of the metabolites across all six individual models, collectively explaining the functional basis for carcinogenicity (**SI Table 5**). Interestingly, about 28.7% of the metabolites were predicted to possess carcinogenic function (probability cutoff ≥ 0.5) **(Figure 2F)**, however, at stringent prediction probabilities cutoff of ≥ 0.8, there were only 46 metabolites **(Figure 2G)**. In order to determine the interrelationship between the individual models in predicting carcinogenic metabolites, we computed and compared the prediction probability vectors in the Eigenspace **(Figure 2H, Figure S3I)**. Our results indicate similarities between the proliferation and anti-apoptotic models, as well as electrophiles and anti-apoptotic properties models in the Eigenspace **(Figure 2H)**. Next, to dissect the molecular pathways in which these predicted carcinogenic metabolites are involved, we projected them on the metabolic pathway space and observed selectivity towards certain metabolic pathways **(Figure 2I)**. Moreover, to gain deeper functional insights, we also performed pathway analysis involving over-representation and enrichment analysis on the predicted carcinogenic metabolites. Functional analysis revealed that these predicted carcinogenic metabolites (probability cutoff ≥ 0.5) are specifically associated with Arachidonic acid metabolism, tyrosine metabolism, xenobiotics, drug metabolism, etc **(Figure 2J, SI Table 6)**. Of note, in the Arachidonic acid metabolism, Metabokiller identified some of the already reported endogenous carcinogenic metabolites, such as Prostaglandins, 5-HETE, and leukotrienes (*36*–*38*). Notably, our unbiased functional analysis also revealed the chemical carcinogenesis ontology comprising 37 previously known carcinogenic metabolites. Some of the potential endogenous carcinogenic metabolites include Melphalan (*39*), Chloroacetaldehyde (CAA) (*40*), N-nitrosodimethylamine (NDMA) (*41*), and others. In a nutshell, by using an orthogonal approach leveraging both the biological and chemical properties of carcinogens, Metabokiller revealed a subset of human metabolites that possess carcinogenic properties.

**Experimental validation of Metabokiller predictions on two previously uncharacterized potential carcinogenic metabolites**

Large-scale *in silico* screening of human metabolomes led to the identification of multiple previously uncharacterized carcinogenic metabolites. Next, to validate the Metabokiller predictions, we selected two previously uncharacterized human metabolites i.e., 4-nitrocatechol (4NC) and 3,4-Dihydroxyphenylacetic acid (DP). Of note, DP is a metabolic intermediate of the tyrosine metabolism pathway while 4NC is involved in aminobenzoate degradation **(Figure S4A)**. Importantly, both the compounds were predicted to be non-carcinogenic by Protox-II and lazar, while vNN predicted 4NC to be non-genotoxic, and DP as genotoxic. Metabokiller predicted DP as a potential carcinogen with a prediction probability of 0.74. Moreover, evaluation of the individual models' probabilities revealed that carcinogenic properties of DP are due to its high genotoxic potential, its electrophilic nature, its ability to alter the epigenetic landscape, and having pro-proliferative and anti-apoptotic properties. Of note, in the case of oxidative stress, we observed below threshold values for the DP, suggesting its inability to induce reactive oxygen species (oxidative stress). Similar to DP, 4NC was also predicted as a potential carcinogen with equivalent prediction probability, however, unlike DP, 4NC qualifies in all the individual models, except for anti-apoptotic properties **(Figure 3A)**. We first investigated the potential of both of these metabolites for their genotoxicity by performing a single-cell gel electrophoresis-based comet assay, a standard method used to measure DNA breaks and lesions (*42*). Of note, these experiments were performed on *Saccharomyces cerevisiae* (budding yeast). Briefly, the yeast was exposed to these metabolites at different concentrations, and post-exposure, the lysed cells were stained using Giemsa dye. We used hydroxyurea (HU), a known DNA repair inhibitor with mutagenic and genotoxic abilities as a positive control in these experiments (*43*). Quantitative assessment of the comet tail lengths revealed a dose-dependent, and the statistically significant increase in the case of 4NC and DP treatment, suggesting their genotoxicity potential **(Figure 3B, C)**. Since we used AMES test results as an input to train the genotoxic model, we next asked whether 4NC and DP also possess mutagenic properties. To address this hypothesis, we performed

mutagenesis testing experiments using the high-throughput fluctuation assay and screened for canavanine-resistant mutants. Of note, we tested different concentrations of 4NC and DP alongside negative vehicle controls **(Figure 3D)**. Our results revealed a significant increase in the number of canavanine mutants in the treated conditions in contrast to the negative control, validating the mutagenic properties of both DP and 4NC (**Figure 3E, F**). Higher concentrations of certain biomolecules often cause pleiotropic effects and indirectly impair the key cellular processes. To test whether the concentrations of the metabolites used for the genotoxic or mutagenesis assays significantly alter the survival and physiology of budding yeast, we performed a propidium iodide-based cell death estimation assay, and observed no profound cell deaths in all the tested concentrations **(Figure 3G, Figure S4B and D)**. Since, 4NC, but not DP is predicted to harbor components of oxidative stress, we next tested the potential of both of these metabolites to induce reactive oxygen species (ROS) levels via fluorometric assay using 2',7'-dichlorofluorescin diacetate (DCFH-DA) dye. Our results indicate a selective increase in the ROS levels in the case of 4NC, but not DP at all tested concentrations, suggesting larger synergy with the prediction probabilities using Metabokiller **(Figure 3H, Figure S4C)**. Since Metabokiller predicted both of these human metabolites as potential carcinogens, we next ask whether the mutations caused by 4NC and DP can alter the genes associated with cell cycle regulation, DNA repair response pathways, or other pathways related to carcinogenesis. To test this, we performed two consecutive selections. The first one involves the selection of the canavanine mutants which in turn confirms the mutagenic effect of the tested metabolites. Canavanine resistant mutants (n=167) were selected and subjected to a second selection for the alterations in the growth assay **(Figure 3I)**. After obtaining the growth profiles of these distinct canavanine-resistant mutants as well as a few randomly selected wild-type clones (omitted for selection 1), we performed hierarchical clustering based on their growth kinetics. Interestingly, we observed three clusters with (a) a few yeast colonies displaying accelerated growth (b) similar growth patterns as that of wild-type yeast, and (c) decelerated growth **(Figure 3J, Figure S4E)**. These results indicate that both 4NC and DP cause random mutagenesis and therefore result in

distinct phenotypes post selections. In summary, all the aforementioned results collectively advocate for the accurate prediction by Metabokiller.

**Metabokiller predicted carcinogenic metabolites undergo mutation of cancer-associated genes and trigger malignancy**

To investigate the impaired molecular mechanisms triggered by DP and 4NC treatment, we performed paired-end Deep RNA sequencing on wild-type and mutant yeast cells in biological triplicates. Notably, only those mutants were selected that were canavanine-resistant (first selection) and displayed accelerated cell division (second selection) **(Figure 4A)**. Differential Gene Expression analysis **(Figure 4D, FigureS5 A-F)** on the normalized read counts identified a comparatively larger number of downregulated genes, both in the case of DP and 4NC with respect to wild-type controls **(Figure 4B)**. Moreover, no overlap was observed in the differentially expressed genes between 4NC and DP vs wild-type, suggesting a difference in their mode-of-action at the molecular level **(Figure 4C)**. We next attempted to decipher the impacted functional pathways impaired in these mutants using Gene Ontology Slim Term Mapper. Our analysis revealed the impairments in genes that are involved in DNA repair, DNA recombination, and cell cycles (**Figure 4E, F, SI Table 7-10**). Among the differentially expressed genes in the case of 4NC and DP, we observed altered expression in genes involved in chromatin organization such as *YOL017W (ESC8)* and *YOR338W* (4NC vs wild type), and *YOL017W (ESC8)* and *YPR018W (RLF2)* (DP vs wild type). Similarly, critical regulators of DNA repair response genes were altered by DP, such as *YDL059C* (*RAD59*) and *YMR173W* (*DDR48*). In addition to these, genes involved in cellular stress response were altered in both cases (**Figure 4E**). In summary, these unbiased analyses further explained the phenotypic changes observed in these mutants using the functional assays. Recent reports showed the possibility of inferring insertion-deletion (indel) mutations using RNA sequencing. We next tracked indel mutations in wild-type and mutants using the raw FASTQ files. Mutational analysis identified specific indel mutations in the *CAN1* gene, validating the selection procedure **(Figure S5G)**. Using strong filtering conditions (supporting read counts ≥ 5), we identified mutations in multiple protein-coding genes that are known to be involved in regulating DNA repair

pathways as well as the cell cycle **(Figure 4G)**. These independent analyses further support and explain the phenotypic response of the mutants (both 4NC and DP). Finally, we tested the carcinogenic effects of DP and 4NC on BEAS-2B normal human epithelial cells using soft-agar assay, a hallmark functional test for detecting malignant transformation. Our results clearly demonstrate that both 4NC and DP can induce malignant transformation of these cells, which collectively reinforce their carcinogenic potential (**Figure 4H-J**). Taken together, a combination of transcriptome and mutational analysis, coupled with malignant transformation assay strongly suggests the carcinogenic role of 4NC and DP via genotoxicity.

## DISCUSSION

The integrity of a cellular genome is always vulnerable to both intrinsic and extrinsic factors (*44*, *45*). A large degree of effort is ongoing to delineate the compositional chemistries of the exogenous chemicals that possess a direct threat to genomic fidelity (*44*). However, recent evidence suggests an equal degree of threat from the endogenous metabolites (*46*). Irrespective of the source, the DNA alterations induced from either source have been implicated in cancer, as well as aging (*47*, *48*). Therefore, accurate evaluation of the carcinogenic potential of any chemical compound is immensely important and plays a vital role in decreasing the cancer burden. Despite the availability of a stringent mandate for carcinogenicity testing, many of the Food and Drug Administration (FDA) approved drugs have been identified as carcinogens and therefore later withdrawn from the market (*49*). As a precautionary measure, stringent protocols have been set up to test the proposed drugs for their carcinogenicity potential. The QSAR models have been used for the computation-assisted pre-screening for carcinogenicity (*17*, *18*, *21*, *24*, *25*). Though multiple computational approaches have been used for the carcinogenicity predictions, unfortunately, they largely fall short on unseen data, primarily due to the limited amount of available data used for model training. In this work, we used an orthogonal approach where, unlike utilizing limited information of *bonafide* carcinogens and non-carcinogens, we leverage chemical and biological properties associated with known carcinogens. Our method, Metabokiller, outperformed most of the recent and widely used methods for carcinogenicity prediction. Moreover, in addition to the higher prediction accuracy, Metabokiller also possesses features of explainable artificial intelligence, since it provides the individual contribution of all the six-core models detailing the biochemical properties towards carcinogenicity **(Figure 5)**. We used Metabokiller to perform a large-scale computational screening of human metabolites for their carcinogenic potential. We identified a large number of previously known and unknown human metabolites with carcinogenicity potential. Our functional analysis of carcinogen-predicted metabolites on pathway space revealed selective enrichment of the Arachidonic acid (AA) metabolism pathway. Interestingly, metabolic intermediates of the AA pathway have been implicated with DNA damage, and are associated with cancer, cardiovascular diseases, as well as

aging (*50*, *51*). Moreover, AA itself is known to cause single-strand DNA breaks as well as oxidative base modifications in the genomic DNA (*52*). Our unbiased screening uncovered other members of the AA pathway that possess carcinogenic potential. Moreover, quantitative assessment of 7,8-dihydro-8-oxo-2‘-deoxyguanosine (8-oxodG) levels in 2’-deoxyguanosine (dG) and isolated DNA revealed the genotoxic potential of AA via oxidative stress (*53*). Other metabolic intermediates of the AA pathway which are known carcinogenic metabolites such as Prostaglandins, 5-HETE, and leukotrienes (*36*–*38*) were also identified using Metabokiller. Further, functional overrepresentation analysis of predicted carcinogenic metabolites identified 37 previously known carcinogenic metabolites. These results strongly advocate for the robustness of biochemical features in capturing carcinogenicity-inducing features.

Despite possessing multiple advantages and immunity towards limited training data for carcinogens and non-carcinogens, Metabokiller also possesses several limitations. First, it has been known that carcinogens also possess other biological properties such as induction of chronic inflammation, inhibition of senescence, cell transformation, changes in growth factors, energetics and signaling pathways related to cellular replication or cell cycle control, angiogenesis, and are immunosuppressive in nature (*54*). The present ensemble model does not take into account these properties. This is mainly due to the paucity of these chemicals in the available literature. Second, an inert compound can also harbor carcinogenic properties after being processed by the cellular enzymes (*55*). These assumptions are also not taken into account in the present ensemble model of Metabokiller. Last, the carcinogenicity of certain molecules is dose-dependent, however, Metabokiller does not take into account the dosage information for predictions. Irrespective of these aforementioned limitations, Metabokiller significantly fills this knowledge gap and provides a quick, cheap but robust, reliable, and accurate alternative for carcinogenicity predictions. Moreover, the list of predicted carcinogenic metabolites opens a new avenue in functional metabolomics and might contribute to unfolding the role of cancer-associated metabolites in disease progression.

## MATERIALS AND METHODS

### Data compilation

In order to build an Ensemble model for carcinogenicity prediction, we opted to utilize the biological and chemical properties associated with carcinogens. These include the potential to induce proliferation response, cellular oxidative stress, epigenetic alterations, genomic instability, and anti-apoptotic properties. Since carcinogens attack DNA by virtue of their electrophilic nature, we, therefore, selected electrophilic potential as one of the parameters. The training data for the six independent models were compiled from the literature or numerous databases in three consecutive steps. Of note, data curation, filtering, and cross-validation were achieved manually to ensure its authenticity. In the first phase, we compiled data from all the available literature, databases, and web servers. In the second phase, we assimilated both the pro and anti or no-activity chemicals for each of the six models from all the aforementioned sources (**SI Table 1**). Finally, in the third phase, we filtered, cross-validated, thoroughly rechecked, and removed the contradictory entries. Post these stringent compilation and screening steps, we gathered 5996 positive and 9613 negative chemicals, that were further segregated based on their aforementioned biological, and chemical features. Taken together, we obtained 6542 chemicals for Genotoxicity model (class 0 = 3026 and class 1 = 3516), 1917 chemicals for electrophilic property model (class 0 = 757 and class 1 = 1160), 753 chemicals for epigenetic modification model (class 0 = 493 and class 1 = 260), 569 chemicals for anti-apoptotic properties model (class 0 = 353 and class 1 = 216), 280 chemicals for proliferative property model (class 0 = 210 and class 1 = 70) and 5548 chemicals for oxidative stress model (class 0 = 4774 and class 1 = 774). In addition to the aforementioned datasets, we have also gathered the list of *bonafide* carcinogens and non-carcinogens from multiple sources and subsequently annotated them as C1, C2, and C3 based on their source information. All these datasets are detailed in Supplementary Tables (**SI Table 2-4**).

### Model Building

Metabokiller Ensemble model was built using six independent predictive models. Each model uses independent training data that is further segregated into two classes based

on their known activity i.e. class 0: anti- or no-activity e.g. anti-genotoxicity and class 1 pro-activity, e.g. pro-genotoxicity. For all the models, the training dataset was evaluated for the class imbalance, and subsequently, the upsampling technique was used to counteract this bias-prone process. Different featurizer and classification models were implemented for each of the models. For instance, in the case of the Epigenetic modification model, we used a graph convolution network (GCN) and MolGraphConv was used as features. In this approach, the node representations in graphs are updated with a variant of GCN. For each graph, its representation is computed by (1) a weighted sum of the node representations in the graph, where the weights are computed by applying a gating function to the node representations (2) a max-pooling of the node representations (3) concatenating the output from point 1 and 2. The final predictions are performed using the Multi-layer Perceptron classifier (MLP). In the case of the genotoxicity prediction model, we used GraphConvModel and the feature generation was performed using ConvMolFeaturizer, an inbuilt function in the DeepChem library (*56*). For the oxidative stress model, we utilized Attentive FP and the feature generation was achieved using MolGraphConv. A similar approach has been used for another model that classifies chemicals based on their electrophilic properties. Last, in the case of the Proliferative and Anti-apoptotic model, we used molecular descriptors-based machine learning techniques. The classifiers used for the Proliferative and anti-apoptotic model are Extra Tree (ET) and Random Forest (RF), respectively. In both of these methods, we used molecular descriptors from the MORDRED software as features (*57*). These two models get re-trained for every new input file. Notably, for the feature selection we used Boruta, a feature selection algorithm, and for the upsampling, Synthetic Minority Oversampling Technique (SMOTE) was used (*58*). For five models, 10 fold cross-validation was performed to estimate the model performance, except in the case of genotoxic properties for which 100 random interactions were performed. The C1 dataset was used as a validation dataset to compute the accuracies for each of the models. We then assigned distinct weights to each of these six models based on their accuracy scores, and finally generated an ensemble model. Finally, we validated the performance of the ensemble model (Metabokiller) on two independent datasets

annotated as C1 and C2, that contains a list of experimentally validated carcinogens and non-carcinogens.

**Strains**

BY4741 strain (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) of *Saccharomyces cerevisiae* was used in all the experiments. Unless mentioned the yeast is grown at 30°C at 200 Revolutions Per Minute (rpm) in the Yeast Extract–Peptone–Dextrose (YPD) media.

**Comet Assay**

Yeast secondary culture with an estimated optical density of 0.6-0.7 is used for the comet assay. Briefly, secondary culture is exposed to different concentrations of 4-nitrocatechol (4NC) (N15553, Sigma-Aldrich), 3,4-Dihydroxyphenylacetic acid (DP) (11569, Sigma-Aldrich) at the final concentrations of 0.1 μM, 1 μM, 10 μM, and 100 μM (n = 3). Moreover, hydroxyurea and solvent alone were used as positive and negative controls. The yeast cells were grown in the presence of metabolite/chemical for 90 minutes at 30°C at 200 rpm. Post-treatment, an equal number of cells were taken from each condition and pelleted down using centrifugation. Pellets were washed twice with pre-cooled 1x Phosphate-buffered saline (PBS) to remove media traces. We resuspended the pellets in 0.3% w/v agarose prepared in S-Buffer (1 M sorbitol, 25 mM KH2PO4, pH 6.5). We next added 40 units of zymolyase enzyme (L2524, Sigma-Aldrich) to degrade the cell wall. Finally, 80 μL of this solution was spread onto glass slides pre-coated with 0.8% w/v agarose. Slides were incubated at 30°C for 30 min for cell wall degradation. Finally, the slides were incubated in the freshly prepared lysis solution (30 mM NaOH, 1 M NaCl, 0.05% w/v SDS, 50 mM EDTA, 10 mM Tris–HCl, pH 10) for 2 hours. The slides were washed three times for 20 min each with electrophoresis buffer (30 mM NaOH, 10 mM EDTA, 10 mM Tris–HCl, pH 10) to remove the lysis solution. We next performed an electrophoresis step in the electrophoresis buffer for 15 min at 70 mV/cm. Post electrophoresis, the slides were incubated in the neutralization buffer (10 mM Tris–HCl, pH 7.4) for 10 mins, followed by 10 minutes incubation in 76% and then 96% ethanol. After this, Giemsa stain was added for 20 min to stain the DNA. The slides were then washed thrice with 1x PBST (0.1% Tween20) for

10 min each. The slides were visualized in a brightfield microscope. Micrographs were taken from random sections of the slides. Approximately 500 cells were chosen randomly from each condition and the comet tail length was measured using ImageJ software.

**Fluctuation Assay**

Yeast was grown overnight at 30°C at 200 rpm in YPD media. An equal number of cells were inoculated into multiple 96 well plates containing 100 µL of YPD media along with respective metabolites (4NC and DP) at concentrations of 0.1 µM, 1 µM, and 10 µM each. The solvent was used as a negative control. A total of 40 biological replicates were set up for each condition. The 96 well plates were covered with aluminum foil to prevent media evaporation. Plates were incubated at 30°C at 200 rpm and the media was replenished with fresh media post 24 hours incubation with aforementioned metabolites or solvents. Post 48 hours of exposure, a Spot assay was performed on Synthetic Complete Medium containing canavanine (30 µg/mL) agar plates, where an equal number of cells were taken from each condition. The plates were incubated at 30°C for 2-3 days, and the number of colonies was counted, and subsequently, Colony Forming Units per mL of media (CFU/ml) was calculated using the following formula:

$$CFU/mL = \frac{\#\ of\ colonies \times Dilution\ Factor}{Volume\ of\ culture\ plated}$$

**Cell Death Assay**

Propidium Iodide staining was used for the quantification of cell deaths in the treated and untreated conditions. Cells were grown overnight at 30°C at 200 rpm in YPD media. Using this primary culture as inoculum, cells were grown in 96 well deep plates in the presence of respective metabolites (4NC and DP) at concentrations of 0.1 µM, 1 µM, and 10 µM each for 12 hours with 8 biological replicates. Propidium iodide (11195, SRL) staining was performed on an aliquot after nine and twelve hours of incubation in 96 well fluorescence plates. Heat-killed cells were used as a positive control. Propidium iodide was added at a final concentration of 5 µg/mL and incubated in the dark for 15

minutes. The fluorescence was measured using Biotek Synergy HTX multi-mode reader at excitation and emission wavelengths of 530/25 and 590/25, respectively.

**Reactive Oxygen Species quantification Assay**

DCFH-DA staining was used to measure the levels of reactive oxygen species. Yeast was grown overnight at 30°C at 200 rpm in YPD media. Using this primary culture as inoculum, cells were grown in 96 well deep plates in the presence of respective metabolites (4NC and DP) at concentrations of 0.1 μM, 1 μM, and 10 μM each for 12 hours with 8 biological replicates. DCFH-DA fluorescence staining was performed after nine and twelve hours of incubation in a 96 well fluorescence plate with 10 mM hydrogen peroxide as a positive control. DCFH-DA (85048, SRL) was added at a final concentration of 10 μM. The plate was incubated in the dark for 30 minutes. The fluorescence was measured using Biotek Synergy HTX multi-mode at excitation and emission wavelengths of 485/20 and 525/20, respectively.

**Selection of Canavanine-resistant mutants with Accelerated Cell Division**

Yeast (BY4741) was grown at 30°C at 200 rpm in YPD media. Using this primary culture as inoculum, cells were grown in MCTs in the presence of respective metabolites (4NC and DP) at concentrations of 0.1 μM, 1 μM, and 10 μM each for 72 hours with 4 biological replicates, alongside negative control. The culture was centrifuged and resuspended in fresh media after every 24 hours. After 72 hours, the cells were plated onto Synthetic Complete Medium containing canavanine (30 μg/mL) agar plates to screen for canavanine-resistant colonies. This is referred to as the first round of selection. The canavanine-resistant colonies from all the conditions were grown in YPD media in the presence of canavanine but without metabolites, and a second selection was performed using the growth assay. RNA was isolated from the canavanine resistant mutants harboring accelerated growth, and post quality check was used for library preparation and sequencing.

**RNA isolation**

Cells (BY4741) were grown at 30°C at 200 rpm in Synthetic Complete Medium containing canavanine (30 µg/mL) in biological triplicates and harvested and resuspended in 500 µL of 1x PBS. Zymolyase (40 units) treatment was performed by incubating for 30 min at 30°C. Post zymolyase treatment, cells were lysed by mechanical breakdown using the glass beads. The lysed cell suspension was transferred to fresh MCT and centrifuged and the pellet was resuspended in 500 µL of TRIzol™ reagent (15596026, Ambion life technologies). 500 µL of chilled chloroform was added to each tube, followed by 15 min of mechanical breakdown and 5 minutes of incubation at room temperature was performed. Samples containing TRIzol™ and chloroform were centrifuged at 12000 g for 15 minutes at 4°C. The upper aqueous layer was carefully transferred to a fresh MCT containing 250 µL of isopropanol. 1 µL of glycogen was also added to accelerate RNA precipitation, followed by transient incubation for 30 minutes at -80°C. Nucleic acid was pelleted down by centrifugation at 12000 g for 15 min at 4°C. Pellet was washed with 70% ethanol at 5000 g for 5 minutes at 4°C and air dried, and subsequently resuspended in RNase-free water. To remove DNA contamination, samples were treated with DNase at 37°C for 15 minutes. DNase inactivation was performed using 5 mM EDTA and heat at 65°C for 10 minutes. This purified RNA was resuspended in RNase-free water for downstream processing.

**RNA sequencing Analysis**

Post sequencing the quality control check was performed using MultiFastQ. The paired-end sequencing read files were mapped to the yeast reference genome (ENSEMBL; Ensembl Fungi release 51) using the align function of the Rsubread package (v2.6.4) (*59*), and therefore, the aligned BAM files were generated. Trimming of reads was performed from both ends prior to mapping. We used the inbuilt function of Rsubread to detect short indel (Insertions and Deletions) mutations and the resulting information was stored in the Variant Call Format (VCF) format. An expression matrix containing the raw read counts for each sample (**SI Table 9**) was obtained using the featureCounts function. Read count matrix was normalized and analyzed using the RUVg function of the RUVSeq package (v1.24.0) (*60*) to remove unwanted variation from RNA-Seq data. Differentially expressed genes were computed using the DESeq2

package (v1.30.1) (*61*). The differentially expressed genes with a fold change cutoff of 1.5, and a P-value significance of < 0.05. Functional Gene Ontology analysis was performed using the GO Slim Mapper tool of SGD (https://www.yeastgenome.org/goSlimMapper) and Metascape (https://metascape.org/gp/index.html).

**Mutation Analysis**

Variant Call Format (VCF) files containing indel mutation information were generated using the Rsubread package as discussed above. We next merged the mutation information from three replicates and a single VCF file was generated using the VCFtools (v0.1.16-1) (*62*). For the downstream analysis, we only selected those mutations that carry the number of supporting reads for variants (SR) ≥ 5. Cross comparison of the VCF files from treatment groups such as 4NC and DP was performed with that of negative control (untreated). Finally, we only selected those variants that were unique to the treatment condition. The Sorting Intolerant From the Tolerant (SIFT) analysis was performed using SIFT 4G Annotator using R64 (sacCer3) as a reference genome (*63*).

**Soft Agar Assay**

BEAS-2B cells were cultured in DMEM media supplemented with 10% FBS and 1% Penicillin-Streptomycin at 37°C humidified incubator with 5% $CO_2$. For soft agar assay, BEAS-2B cell suspension of $1\times10^4$ cells/well was added to top agar solution at a final concentration of 0.3% and final volume of 0.5 ml per well and suspended on the top of solidified 0.5% agar base. The overnight culture was treated with 10 μM DP or 10 μM 4NC. The cells were supplemented with Dulbecco's Modified Eagle Medium (DMEM) complete media twice until colonies appeared after 3 weeks. Colonies were stained with 0.1% crystal violet in 10% ethanol and imaged using a 4x lens and bright-field filter in the light microscope.

**Statistical Analysis**

All statistical analyses were performed using the Past 4. The P-value cut-off used in this study is 0.05. *, **, ***, and **** refer to P-values <0.05, <0.01, <0.001, and <0.0001, respectively. For comparison of the medians of the two distributions, Mann–Whitney U test was used, while a two-tailed student's t-test was used for comparing the means. A negative Binomial test was used to measure significance in the Fluctuation Assay.

## FIGURE LEGENDS

**Figure 1: Metabokiller, an explainable artificial intelligence-driven ensemble model for carcinogen prediction.**

**(A)** Schematic representation depicting the key steps involved in building Metabokiller. Of note, six independent classification models that form the core of Metabokiller prediction engines are highlighted in different colors. Weights are assigned to each of these six models based on their accuracy in predicting *bonafide* carcinogens and non-carcinogens. **(B)** Bar graphs depicting the number of chemicals in the indicated conditions that were used as an input in building these six independent prediction models. Of note, grey and black bars indicate the number of chemicals assigned as a pro and anti/No activities of the indicated biochemical processes. **(C)** Three-dimensional Principal Component Analysis reveals the chemical heterogeneity between the two classes (pro and anti/No) of the indicated datasets. Heatmaps at the bottom depict the relative enrichment of the indicated functional groups in both classes. **(D)** AUC (Area under the curve) plots represent the performance of indicated best models on the held-out data. **(E)** Box plots depicting the accuracy of the 10 fold cross-validation of the indicated models. Of note, in the case of the Genotoxicity dataset (annotated with asterisk sign), a box plot displaying the results of random 100 iterations. **(F)** Box plots depicting the AUCROC of the 10 fold cross-validation of the indicated models. Of note, in the case of the Genotoxicity dataset (annotated with asterisk sign), a box plot displaying the results of random 100 iterations. **(G)** Bar plots depicting the weights assigned to each of the indicated models based on their performance (accuracy) on the held-out testing (C1) dataset.

**Figure 2: Metabokiller revealed human metabolites possessing carcinogenic potential, and outperformed other prediction methods.**

**(A)** Schematic representation of the projection of two external datasets, named C2 and C3 that harbor information about *bonafide* carcinogens and non-carcinogens, on Metabokiller. **(B)** AUC (Area under the curve) plots represent the performance of Metabokiller on two external carcinogen datasets (C2 and C3 datasets). **(C)** Diagrammatic representation depicting the experimental workflow to cross-compare the performance of Metabokiller alongside other widely accepted methods for carcinogen predictions. Notably, the C2 dataset was prefiltered for the removal of shared data points that were used in the training data for any of the four models to avoid bias. **(D)** Bar plots depicting the performance of the indicated models on unseen C2 dataset. **(E)** Schematic representation depicting the projection of The Human Metabolome Database (HMDB) on Metabokiller Ensemble model for the prediction of carcinogenic and non-carcinogenic metabolites. **(F)** Overlapping density plots depicting the distribution of computed prediction probabilities of human metabolites using Metabokiller. Of note, the red and green color represents the predicted output from Metabokiller. **(G)** Venn diagram depicting the total number of human metabolites used as an input in Metabokiller for carcinogenicity prediction. Of note, the number of metabolites was further segregated based on different prediction probability cutoffs. **(H)** Variables factor map (PCA) depicting all the six variables (individual models) of the predicted carcinogens of HMDB in the PCA dimensionality space. **(I)** Projection of the predicted carcinogens (indicated as red dots) on the human metabolic space, achieved using iPath Web Server. **(J)** Circle packing chart representing metabolic pathway enrichment results of the predicted carcinogenic metabolites using Metabokiller. Of note, the pathway information is gathered from the BioCarta (blue), HumanCyc (orange), KEGG (green), and SMPDB (red), and the analysis is performed using Integrated Molecular Pathway Level Analysis (IMPaLA).

**Figure 3: High-throughput assays experimentally validates the Metabokiller predictions**

**(A)** Bar graphs denoting the prediction probabilities of the indicated six sub-models of Metabokiller for the two previously uncharacterized human metabolites i.e. 4-Nitrocatechol (4NC), and 3,4-Dihydroxyphenylacetic acid (DP). **(B)** Representative micrographs depicting the results of comet assay performed on budding yeast in the indicated conditions. Of note, NC and hydroxyurea (HU) represent the negative and positive controls, respectively. **(C)** Violin plot depicting the comparative distribution of the comet tail length of ~500 randomly selected cells per condition. Of note, four different concentrations of 4-Nitrocatechol and 3,4-Dihydroxyphenylacetic acid were used. Mann–Whitney U test was used to compute statistical significance between the test conditions and the negative control. **(D)** Schematic diagram depicting the experimental workflow in a 96 well plate format for testing the mutagenic potential of the 4-Nitrocatechol and 3,4-Dihydroxyphenylacetic acid. Of note, three different concentrations of 4-Nitrocatechol and 3,4-Dihydroxyphenylacetic acid were used. **(E)** Representative micrographs depicting the mutagenesis effect of 4-Nitrocatechol and 3,4-Dihydroxyphenylacetic acid in conferring canavanine resistance mutation identified using spot assay. **(F)** Bar plots depicting the colony-forming units (CFU) per ml in the indicated conditions. Of note, the Negative Binomial test was used to compute statistical significance between the test conditions and the negative control. **(G)** Bar plots depicting the normalized fluorescence of propidium iodide staining from the cell death assay in the indicated conditions. Of note, heat-killed (HK) yeast is used as a positive control. The student's t-test was used to compute statistical significance between the test conditions and the negative control (NC). **(H)** Box plot depicting the results of reactive oxygen species (ROS) levels inferred using DCHF dye-based assay in the indicated conditions. Mann–Whitney U test was used to compute statistical significance between the test conditions and the negative control (NC). Notably, hydrogen peroxide ($H_2O_2$) treated yeast cells were used as a positive control. **(I)** Diagrammatic representation of the experimental design to select the gain of function mutants for accelerated cell division. Of note, mutants are selected based on canavanine resistance (selection 1) as well as accelerated cell division (selection 2). **(J)** Growth curve profiles

of untreated wild-type (NC) and 4-Nitrocatechol (4NC) and 3,4-Dihydroxyphenylacetic acid (DP) induced mutants with accelerated growth under optimal growth conditions. The student's t-test was used to compute statistical significance between the test conditions and the NC.

**Figure 4: Carcinogenic and mutagenic potential of 4-Nitrocatechol (4NC) and 3,4-Dihydroxyphenylacetic acid (DP) confirmed using Deep RNA Sequencing and Soft Agar Assay.**

**(A)** Schematic representation depicting the RNA sequencing experimental design, featuring the group information, treatment duration, and the sequencing parameters. **(B)** Bar graph depicting the number of differentially expressed genes between the untreated and treatment conditions (4-Nitrocatechol (4NC) and 3,4-Dihydroxyphenylacetic acid (DP)). The y-axis represents the number of statistically significant genes in the indicated conditions, whereas the x-axis represents the treatment group information. **(C)** Venn diagram depicting the number of differentially expressed genes shared among 4-Nitrocatechol (4NC) and 3,4-Dihydroxyphenylacetic acid (DP). **(D)** Heatmap representation of the differentially expressed genes in DP and 4NC with respect to untreated conditions. **(E)** Gene Ontology Slim Mapper analysis depicting the association of the differentially expressed genes in the indicated ontologies. **(F)** Metascape-based Functional Gene Ontology analysis identified the involvement of differentially expressed genes in the indicated prominent biological processes. **(G)** The Venn diagram and its accompanying heatmap on the right side collectively depict the number of mutated genes (Insertion Deletion mutations) involved in DNA repair or Cell Cycle regulation. Heatmap on the right further segregates the identified mutations as frameshift (insertion or deletion) and substitution. **(H)** Representative micrographs depicting the results of malignant transformation assay. Scale bar = 100 μm. **(I)** Mean-Whisker plot depicting the distribution of the malignantly transformed colonies in the indicated conditions. Mann Witney's pairwise test was used to compute the statistical significance, with a p-value cutoff of $< 0.05$. **(J)** Bar graphs represent the fold change increase in the number of transformed colonies in the indicated conditions. Mann Witney's pairwise test was used to compute the statistical significance, with a p-value cutoff of $< 0.05$.

**Figure 5: Graphical Abstract of the key findings**

Schematic representation depicting the functionality of the Metabokiller prediction engine. Of note, Metabokiller predicts the carcinogenicity potential of any query chemical by assessing for the indicated biochemical properties.

## SUPPLEMENTARY FIGURE LEGENDS

**Figure S1: Metabokiller: An ensemble model that leverages biochemical properties to predict carcinogenicity.**

**(A)** Diagram depicting the step-by-step workflow used to build all the individual six models as well as the ensemble model. Of note, six independent training datasets were used to build six independent models. The class imbalance problem was resolved using the upsampling approach. Distinct methods were used to compute the numerical features from the input chemicals from all six datasets. Hyperparameter tuning was performed for the identification of best-performing parameters. An external carcinogen dataset was subsequently used to assign weights to each of the six aforementioned models to build the ensemble model. **(B)** Workflow diagram depicting the strategy employed to test the external datasets on Metabokiller.

**Figure S2: Metabokiller leverages a blend of molecular descriptors-based machine learning and Graph-based Deep learning approaches**

**(A)** Box plots depicting the classification performances of the indicated models as inferred from the 10 fold cross-validation. Of note, only in the case of the Genotoxicity model, box plot represents the results of random 100 iterations. **(B)** Table representing the details such as the underlying model/classifier and featurizer used for each of the indicated models. **(C)** Tables depicting information about the total number of chemicals, further segregated into two classes and the hyperparameters used to build the indicated models.

**Figure S3: Metabokiller introduce a novel approach for carcinogenicity prediction and outperforms existing methods**

**(A)** Principal Component Analysis in the three-dimensional space revealing the chemical heterogeneity between the two classes (carcinogens and non-carcinogens) in the C2 dataset harboring *bonafide* carcinogens and non-carcinogens. The heatmap at the bottom depicts the relative enrichment of the indicated functional groups in both classes. **(B)** Principal Component Analysis in the three-dimensional space revealing the chemical heterogeneity between the two classes (carcinogens and non-carcinogens) in

the C3 dataset harboring *bonafide* carcinogens and non-carcinogens. Heatmap at the bottom depicts the relative enrichment of the indicated functional groups in both classes **(C)** Bar plots depicting the Metabokiller performance on the external C2 carcinogens/non-carcinogens dataset. **(D)** Bar plots depicting the Metabokiller performance on the external C3 carcinogens/non-carcinogens dataset. **(E)** 3D PCA and confusion matrix depicting the Metabokiller predictions on the C2 independent dataset, segregated into the true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). **(F)** 3D PCA and confusion matrix depicting the Metabokiller predictions on the C3 independent dataset, segregated into true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). **(G)** Venn diagram depicting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) predictions on C2 dataset by the indicated models. **(H)** Table describing the details about other studies used for the carcinogenicity or mutagenicity predictions. **(I)** Variables factor map (PCA) depicting all the six variables (individual models) as well as ensemble predictions of the HMDB metabolites in the PCA dimensionality space.

**Figure S4: High synergy between Metabokiller predictions and experimental validations.**

**(A)** Schematic representation depicting the metabolic intermediates harboring predicted carcinogens and non-carcinogens in the tyrosine metabolism pathway as well as aminobenzoate degradation pathway. **(B)** Box plots depicting the fluorescence intensity of propidium iodide staining from the cell death assay in the indicated conditions. Of note, heat-killed (HK) yeast is used as a positive control. Mann–Whitney U test was used to compute statistical significance between the test conditions and the negative control. **(C)** Box plot depicting the results of reactive oxygen species (ROS) levels inferred using DCHF dye-based assay in the indicated conditions. Of note, ROS levels were estimated 12 hours post-incubation with the indicated conditions. Mann–Whitney U test was used to compute statistical significance between the test conditions and the negative control. Notably, hydrogen peroxide ($H_2O_2$) treated yeast cells were used as a positive control. **(D)** Growth curve profiles of the treated and untreated wild-type yeast during transient exposure of the indicated conditions. The Student's t-test was used to

compute statistical significance between the test conditions and the negative control. **(E)** Heatmap depicting the growth profiles of all the metabolite-induced canavanine-resistant mutants. Mutants were clustered based on their growth profiles. Three major clusters were obtained with either accelerated, decelerated, or normal growth. The black rectangle in the middle indicates the growth profiles of the untreated cells.

**Figure S5: RNA sequencing independently validated the functional and mutagenic effect of 4-Nitrocatechol and 3,4-Dihydroxyphenylacetic acid.**
**(A)** Bar plots depicting the total read counts (in millions) of the indicated RNA sequencing samples. **(B)** Box plot representing the distribution of the transformed read count data in the indicated conditions. **(C)** Correlation plot depicting the relationship between the individual RNA sequencing samples. Of note, 75% of the normalized and transformed data is used for the correlation analysis. **(D-E)** Box plots depicting the relative log expression of the individual replicates of the indicated conditions before and after upper quantile normalization. **(F)** Volcano plot indicating the differentially expressed genes between the treated (metabolite treatment) and untreated conditions. **(G)** Genomic representation of the *CAN1* gene loci indicates the mutants-specific insertions or deletions.

## SUPPLEMENTARY INFORMATION

**SI Table 1:** Tabular representation of the sources used to curate the input datasets for the individual six models of the Metabokiller. Of note, for each of the six models, the source information is provided either in the form of a Digital Object Identifier or PubMed ID.

**SI Table 2:** Table containing information about the *bonafide* list of carcinogens and non-carcinogens that collectively constitutes the C1 dataset. Of note, the table contains information about the chemical name, its carcinogenic status, its Simplified molecular-input line-entry system (SMILES), and the source of the data points.

**SI Table 3:** Table containing information about the *bonafide* list of carcinogens and non-carcinogens that collectively constitutes the C2 dataset. Of note, the table contains information about the chemical name, its carcinogenic status, its Simplified molecular-input line-entry system (SMILES), and the source of the data points.

**SI Table 4:** Table containing information about the bonafide list of carcinogens and non-carcinogens that collectively constitutes the C3 dataset. Of note, the table contains information about the chemical name, its carcinogenic status, its Simplified molecular-input line-entry system (SMILES), and the source of the data points.

**SI Table 5:** Table containing information about the prediction results of the Human Metabolome Database using Metabokiller. Of note, the table contains information about chemical names, HMDB compound ID, prediction probabilities as well as their final status at 0.5 cutoffs for all the individual six models. It also contains predictions of the ensemble model that denotes the carcinogenic potential of the listed metabolites.

**SI Table 6:** Table containing information about the results of the functional enrichment of the carcinogen-predicted human metabolites. The results are obtained using IMPaLA: Integrated Molecular Pathway Level Analysis web server. The table contains information

about the pathway names, their source, number of overlapping metabolites, identification details of these metabolites, and the P and Q values.

**SI Table 7:** Table contains information about the Gene Ontology results of the differentially expressed genes upon 4-nitrocatechol (4NC) treatment. The table contains information about Gene Ontology IDs, Term names, and the information of the mapped gene for each of the Gene Ontology. The analysis was performed using the GO Slim Mapper tool from the Saccharomyces Genome Database (SGD).

**SI Table 8:** Table contains information about the Gene Ontology results of the differentially expressed genes upon 3,4-Dihydroxyphenylacetic acid (DP) treatment. The table contains information about Gene Ontology IDs, Term names, and the information of the mapped gene for each of the Gene Ontology. The analysis was performed using the GO Slim Mapper tool from the Saccharomyces Genome Database (SGD).

**SI Table 9:** Table contains information about the unnormalized read count data of the untreated (NC), 3,4-Dihydroxyphenylacetic acid (DP) treated and 4-nitrocatechol (4NC) treated canavanine mutants.

**SI Table 10:** Table contains information about the Trimmed Mean of the M-values (TMM)-normalized read count data of the untreated (NC), 3,4-Dihydroxyphenylacetic acid (DP) treated and 4-nitrocatechol (4NC) treated canavanine mutants.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable

## CONSENT FOR PUBLICATION

All authors have read and approved the manuscript for publication.

## AVAILABILITY OF DATA AND MATERIALS

An end-to-end workflow for the entire analysis is provided in GitHub: https://github.com/the-ahuja-lab/Metabokiller and Zenodo (10.5281/zenodo.5704813). The raw sequencing files are available at ArrayExpress with Accession ID: E-MTAB-11179. The web server for the Metabokiller is available at https://metabokiller.ahujalab.iiitd.edu.in/.

## COMPETING INTERESTS

The authors declare that they have no competing interests.


## FUNDING

The Ahuja lab is supported by the Ramalingaswami Re-entry Fellowship (BT/HRD/35/02/2006), a re-entry scheme of the Department of Biotechnology (DBT), Ministry of Science & Technology, Govt. of India, Start-Up Research Grant (SRG/2020/000232) from the SERB, Science, and Engineering Research Board and an intramural Start-up grant from Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi). The Sengupta lab is funded by the INSPIRE faculty grant from the Department of Science & Technology (DST), India.


## AUTHORS' CONTRIBUTIONS

The study was conceived by G.A. Computational analysis workflows were designed by G.A., and D.S. Yeast and Human experimental workflows were designed by G.A., and S.N., respectively. Yeast-based assays were performed by A.M., S.A., and N.D. Human cell culture-based experiments were performed by S.S.

Data compilation for the model building was performed by P.G., A.A. P.R., A.M., V.G., S.M., and analysis workflow and the webserver was made by R.S., A.M., and P.G. V.S., An.M. and J.T assisted in data interpretation. Metabokiller web server is created by R.S., and R.G. Illustrations were drafted by A.M., V.G., and G.A., and G.A. wrote the paper. All authors have read and approved the manuscript.

**ACKNOWLEDGEMENTS**

The authors would like to thank the IT-HelpDesk team of IIIT-Delhi for providing assistance with the computational resources. We thank all the members of the Ahuja lab for their intellectual contributions at various stages of this project. We thank Dr. Kaustuv Datta for providing critical insights into this study.

## REFERENCES


1. S. M. Rappaport, Redefining environmental exposure for disease etiology. *npj Systems Biology and Applications*. **4**, 1–6 (2018).

2. W. H. Farland, A. Lynch, N. K. Erraguntla, L. H. Pottenger, Improving risk assessment approaches for chemicals with both endogenous and exogenous exposures. *Regul. Toxicol. Pharmacol.* **103**, 210–215 (2019).

3. J. A. Swenberg, K. Lu, B. C. Moeller, L. Gao, P. B. Upton, J. Nakamura, T. B. Starr, Endogenous versus exogenous DNA adducts: their role in carcinogenesis, epidemiology, and risk assessment. *Toxicol. Sci.* **120 Suppl 1**, S130–45 (2011).

4. A. Luch, Nature and nurture - lessons from chemical carcinogenesis. *Nat. Rev. Cancer*. **5**, 113–125 (2005).

5. H. Yasaei, E. Gilham, J. C. Pickles, T. P. Roberts, M. O'Donovan, R. F. Newbold, Carcinogen-specific mutational and epigenetic alterations in INK4A, INK4B and p53 tumour-suppressor genes drive induced senescence bypass in normal diploid mammalian cells. *Oncogene*. **32**, 171–179 (2012).

6. R. P. P. Fuchs, N. Schwartz, M. P. Daune, Hot spots of frameshift mutations induced by the ultimate carcinogen N- acetoxy-N-2-acetylaminofluorene. *Nature*. **294**, 657–659 (1981).

7. L. J. Lilly, B. Bahner, P. N. Magee, Chromosome aberrations induced in rat lymphocytes by N-nitroso compounds as a possible basis for carcinogen screening. *Nature*. **258**, 611–612 (1975).

8. F. Madia, A. Worth, M. Whelan, R. Corvi, Carcinogenicity assessment: Addressing the challenges of cancer and chemicals in the environment. *Environ. Int.* **128**, 417–429 (2019).

9. P. Anand, A. B. Kunnumakkara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, B. B. Aggarwal, Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.* **25**, 2097–2116 (2008).

10. G. M. Williams, M. J. Iatropoulos, J. H. Weisburger, Chemical carcinogen mechanisms of action and implications for testing methodology. *Exp. Toxicol. Pathol.* **48**, 101–111 (1996).

11. J. C. Barrett, Mechanisms of action of known human carcinogens. *IARC Sci. Publ.*, 115–134 (1992).

12. Y. Sun, Z. Shi, H. Lian, P. Cai, Energy metabolic dysfunction as a carcinogenic factor in cancer cells. *Clin. Transl. Med.* **5**, 14 (2016).

13. K. Hardonnière, E. Saunier, A. Lemarié, M. Fernier, I. Gallais, C. Héliès-Toussaint,

B. Mograbi, S. Antonio, P. Bénit, P. Rustin, M. Janin, F. Habarou, C. Ottolenghi, M.-T. Lavault, C. Benelli, O. Sergent, L. Huc, S. Bortoli, D. Lagadic-Gossmann, The environmental carcinogen benzo[a]pyrene induces a Warburg-like metabolic reprogramming dependent on NHE1 and associated with cell survival. *Sci. Rep.* **6**, 1–13 (2016).

14. A. Moretton, J. I. Loizou, Interplay between Cellular Metabolism and the DNA Damage Response in Cancer. *Cancers* . **12** (2020), doi:10.3390/cancers12082051.

15. K. A. Meister, *America's War on "carcinogens": Reassessing the Use of Animal Tests to Predict Human Cancer Risk* (Am Cncl on Science, Health, 2005).

16. M. D. Waters, M. Jackson, I. Lea, Characterizing and predicting carcinogenicity and mode of action using conventional and toxicogenomics methods. *Mutat. Res.* **705**, 184–200 (2010).

17. N. Fjodorova, M. Vračko, M. Tušar, A. Jezierska, M. Novič, R. Kühne, G. Schüürmann, Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses. *Mol. Divers.* **14**, 581–594 (2010).

18. A. H. Morales, M. A. C. Pérez, R. D. Combes, M. P. González, Quantitative structure activity relationship for the computational prediction of nitrocompounds carcinogenicity. *Toxicology*. **220**, 51–62 (2006).

19. R. Benigni, A. Giuliani, R. Franke, A. Gruska, Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chem. Rev.* **100**, 3697–3714 (2000).

20. K. P. Singh, S. Gupta, P. Rai, Predicting carcinogenicity of diverse chemicals using probabilistic neural network modeling approaches. *Toxicol. Appl. Pharmacol.* **272**, 465–475 (2013).

21. L. Zhang, H. Ai, W. Chen, Z. Yin, H. Hu, J. Zhu, J. Zhao, Q. Zhao, H. Liu, CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Sci. Rep.* **7**, 2118 (2017).

22. R. Benigni, Others, The Benigni/Bossa rulebase for mutagenicity and carcinogenicity--a module of Toxtree. *JRC Scientific and Technical Reports*. **1**, 63 (2008).

23. R. Benigni, C. Bossa, O. Tcheremenskaia, A. Giuliani, Alternatives to the carcinogenicity bioassay: in silico methods, and the in vitro and in vivo mutagenicity assays. *Expert Opin. Drug Metab. Toxicol.* **6**, 809–819 (2010).

24. Z. Zhou, Q. Dai, T. Gu, A QSAR model of PAHs carcinogenesis based on thermodynamic stabilities of biactive sites. *J. Chem. Inf. Comput. Sci.* **43**, 615–621 (2003).

25. P. Ruiz, O. Faroon, C. J. Moudgal, H. Hansen, C. T. De Rosa, M. Mumtaz, Prediction of the health effects of polychlorinated biphenyls (PCBs) and their metabolites using quantitative structure-activity relationship (QSAR). *Toxicol. Lett.* **181**, 53–65 (2008).

26. Z. Nozhat, M. S. Khalaji, M. Hedayati, S. K. Kia, Different Methods for Cell Viability and Proliferation Assay: Essential Tools in Pharmaceutical Studies. *Anticancer Agents Med. Chem.* (2020), doi:10.2174/1871520621999201230202614.

27. N. Kumar, R. Afjei, T. F. Massoud, R. Paulmurugan, Comparison of cell-based assays to quantify treatment effects of anticancer drugs identifies a new application for Bodipy-L-cystine to measure apoptosis. *Sci. Rep.* **8**, 16363 (2018).

28. A. M. Tabish, K. Poels, P. Hoet, L. Godderis, Epigenetic factors in cancer risk: effect of chemical carcinogens on global DNA methylation pattern in human TK6 cells. *PLoS One*. **7**, e34674 (2012).

29. E. C. Miller, J. A. Miller, Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer*. **47**, 2327–2345 (1981).

30. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: the next generation. *Cell*. **144**, 646–674 (2011).

31. P. Banerjee, A. O. Eckert, A. K. Schrey, R. Preissner, ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res.* **46**, W257–W263 (2018).

32. P. Schyman, R. Liu, V. Desai, A. Wallqvist, vNN Web Server for ADMET Predictions. *Front. Pharmacol.* **8**, 889 (2017).

33. C. Helma, Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. *Mol. Divers.* **10**, 147–158 (2006).

34. in *Predictive Toxicology* (CRC Press, 2005), pp. 491–512.

35. D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).

36. Á. Jara-Gutiérrez, V. Baladrón, The Role of Prostaglandins in Different Types of Cancer. *Cells*. **10** (2021), doi:10.3390/cells10061487.

37. Z. Wen, H. Liu, M. Li, B. Li, W. Gao, Q. Shao, B. Fan, F. Zhao, Q. Wang, Q. Xie, Y. Yang, J. Yu, X. Qu, Increased metabolites of 5-lipoxygenase from hypoxic ovarian cancer cells promote tumor-associated macrophage infiltration. *Oncogene*. **34**, 1241–1252 (2015).

38. T. Kawamori, N. Uchiya, T. Sugimura, K. Wakabayashi, Enhancement of colon carcinogenesis by prostaglandin E2 administration. *Carcinogenesis*. **24**, 985–990 (2003).

39. Q. Chen, P. C. Van der Sluis, D. Boulware, L. A. Hazlehurst, W. S. Dalton, The FA/BRCA pathway is involved in melphalan-induced DNA interstrand cross-link repair and accounts for melphalan resistance in multiple myeloma cells. *Blood*. **106**, 698–705 (2005).

40. T. Matsuda, T. Yagi, M. Kawanishi, S. Matsui, H. Takebe, Molecular analysis of mutations induced by 2-chloroacetaldehyde, the ultimate carcinogenic form of vinyl chloride, in human cells using shuttle vectors. *Carcinogenesis*. **16**, 2389–2394 (1995).

41. R. H. Adamson, B. A. Chabner, The Finding of N-Nitrosodimethylamine in Common Medicines. *Oncologist*. **25**, 460–462 (2020).

42. F. Azevedo, F. Marques, H. Fokt, R. Oliveira, B. Johansson, Measuring oxidative DNA damage and DNA repair using the yeast comet assay. *Yeast*. **28**, 55–61 (2011).

43. J. L. Santos, P. L. Bosquesi, A. E. Almeida, C. M. Chin, E. A. Varanda, Mutagenic and genotoxic effect of hydroxyurea. *Int. J. Biomed. Sci.* **7**, 263–267 (2011).

44. P. Irigaray, D. Belpomme, Basic properties and molecular mechanisms of exogenous chemical carcinogens. *Carcinogenesis*. **31**, 135–148 (2010).

45. W. K. Lutz, T. Fekete, Endogenous and exogenous factors in carcinogenesis: limits to cancer prevention. *Int. Arch. Occup. Environ. Health*. **68**, 120–125 (1996).

46. N. Rehmani, S. M. Hadi, DNA Reactive Activities of Some Endogenous Metabolites and their Putative Role in the Induction of Cancer, doi:10.4172/1948-5956.1000349.

47. J. H. J. Hoeijmakers, DNA damage, aging, and cancer. *N. Engl. J. Med.* **361**, 1475–1485 (2009).

48. G. Ahuja, D. Bartsch, W. Yao, S. Geissen, S. Frank, A. Aguirre, N. Russ, J.-E. Messling, J. Dodzian, K. A. Lagerborg, N. E. Vargas, J. S. Muck, S. Brodesser, S. Baldus, A. Sachinidis, J. Hescheler, C. Dieterich, A. Trifunovic, A. Papantonis, M. Petrascheck, A. Klinke, M. Jain, D. R. Valenzano, L. Kurian, Loss of genomic integrity induced by lysosphingolipid imbalance drives ageing in the heart. *EMBO Rep.* **20** (2019), doi:10.15252/embr.201847407.

49. V. B. Siramshetty, J. Nickel, C. Omieczynski, B.-O. Gohlke, M. N. Drwal, R. Preissner, WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.* **44**, D1080–D1086 (2016).

50. J. Booyens, C. F. Van der Merwe, I. E. Katzeff, Chronic arachidonic acid eicosanoid imbalance: A common feature in coronary artery disease, hypercholesterolemia, cancer and other important diseases …. *Med. Hypotheses* (1985) (available at https://www.sciencedirect.com/science/article/pii/0306987785901203).

51. Z. Zheng, Y. Li, G. Jin, T. Huang, M. Zou, S. Duan, The biological role of arachidonic acid 12-lipoxygenase (ALOX12) in various human diseases. *Biomed. Pharmacother.* **129**, 110354 (2020).

52. W. J. Kozumbo, N. M. Hanley, S. Agarwal, M. J. Thomas, M. C. Madden, Products of ozonized arachidonic acid potentiate the formation of DNA single strand breaks in cultured human lung cells. *Environ. Mol. Mutagen.* **27**, 185–195 (1996).

53. T. M. de Kok, F. ten Vaarwerk, I. Zwingman, J. M. van Maanen, J. C. Kleinjans, Peroxidation of linoleic, arachidonic and oleic acid in relation to the induction of oxidative DNA damage and cytogenetic effects. *Carcinogenesis*. **15**, 1399–1404 (1994).

54. M. T. Smith, K. Z. Guyton, C. F. Gibbons, J. M. Fritz, C. J. Portier, I. Rusyn, D. M. DeMarini, J. C. Caldwell, R. J. Kavlock, P. F. Lambert, S. S. Hecht, J. R. Bucher, B. W. Stewart, R. A. Baan, V. J. Cogliano, K. Straif, Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ. Health Perspect.* **124**, 713–721 (2016).

55. F. H. Schmidt, A new way to understand chemical carcinogenesis and cancer prevention. *RRMC*. **4**, 23–33 (2014).

56. B. Ramsundar, P. Eastman, E. Feinberg, J. Gomes, K. Leswing, A. Pappu, M. Wu, V. Pande, DeepChem: Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry. *Materials Science and Biology. https://github. com/deepchem/deepchem (accessed Aug 8, 2017)* (2019).

57. H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).

58. R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*. **14**, 106 (2013).

59. Y. Liao, G. K. Smyth, W. Shi, The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).

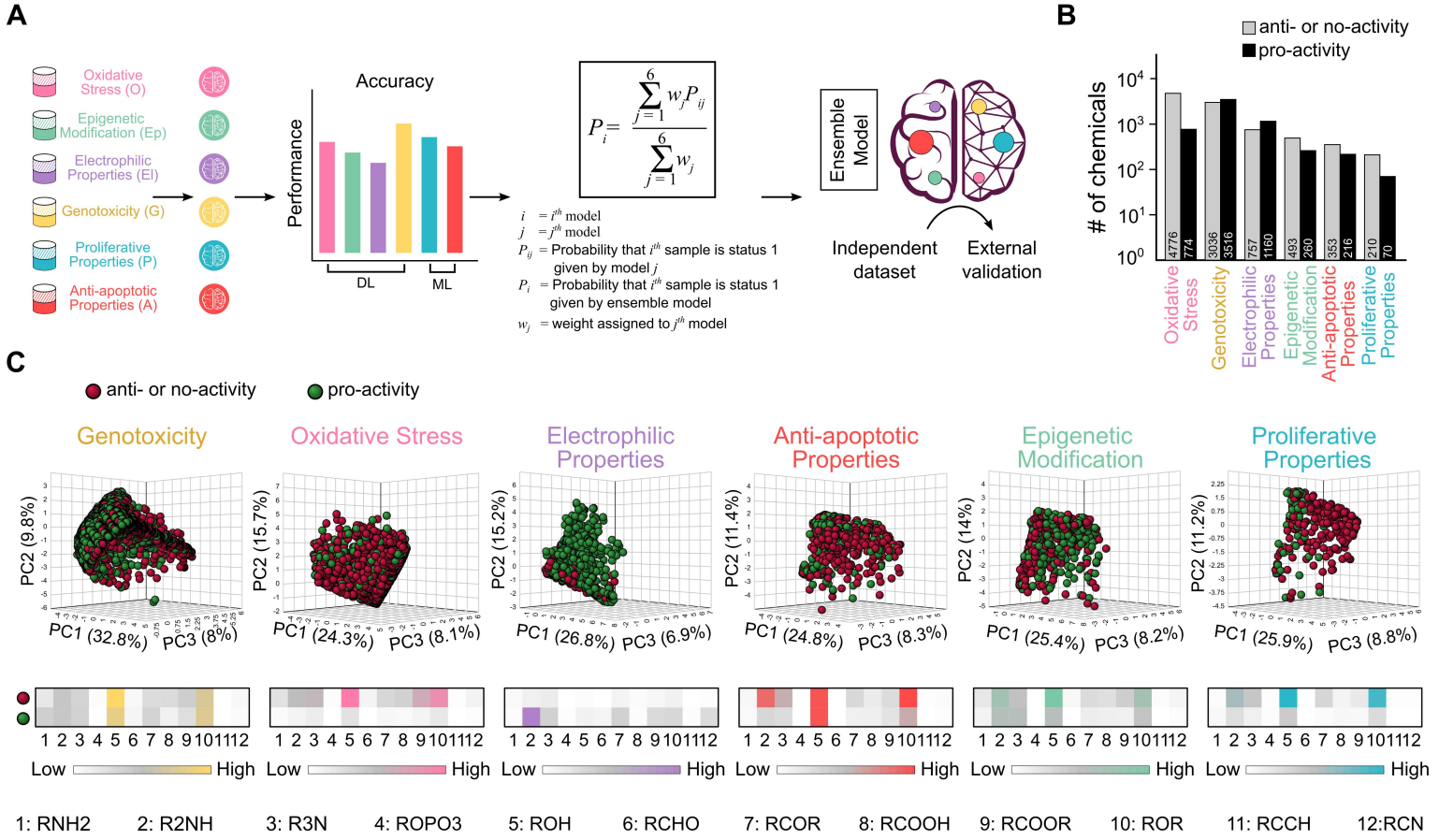
60. D. Risso, RUVSeq: remove unwanted variation from RNA-seq data. *Bioconductor https://bioconductor. org/packages/release/bioc/html/RUVSeq. html* (2015) (available at http://bioconductor.jp/packages/3.1/bioc/vignettes/RUVSeq/inst/doc/RUVSeq.pdf).

61. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and

dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

62. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics*. **27**, 2156–2158 (2011).

63. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).


A
Oxidative Stress (O)
Epigenetic Modification (Ep)
Electrophilic Properties (El)
Genotoxicity (G)
Proliferative Properties (P)
Anti-apoptotic Properties (A)
Accuracy
Performance
DL
ML
$P_i = \frac{\sum_{j=1}^{6} w_j P_{ij}}{\sum_{j=1}^{6} w_j}$
$i$ = $i^{th}$ model
$j$ = $j^{th}$ model
$P_{ij}$ = Probability that $i^{th}$ sample is status 1 given by model $j$
$P_i$ = Probability that $i^{th}$ sample is status 1 given by ensemble model
$w_j$ = weight assigned to $j^{th}$ model
Ensemble Model
Independent dataset
External validation
B
anti- or no-activity
pro-activity
# of chemicals
4776
774
3036
3516
757
1160
493
260
353
216
210
70
Oxidative Stress
Genotoxicity
Electrophilic Properties
Epigenetic Modification
Anti-apoptotic Properties
Proliferative Properties
C
anti- or no-activity
pro-activity
Genotoxicity
Oxidative Stress
Electrophilic Properties
Anti-apoptotic Properties
Epigenetic Modification
Proliferative Properties
PC2 (9.8%)
PC1 (32.8%)
PC3 (8%)
PC2 (15.7%)
PC1 (24.3%)
PC3 (8.1%)
PC2 (15.2%)
PC1 (26.8%)
PC3 (6.9%)
PC2 (11.4%)
PC1 (24.8%)
PC3 (8.3%)
PC2 (14%)
PC1 (25.4%)
PC3 (8.2%)
PC2 (11.2%)
PC1 (25.9%)
PC3 (8.8%)
1 2 3 4 5 6 7 8 9 10 11 12
Low
High
1: RNH2
2: R2NH
3: R3N
4: ROPO3
5: ROH
6: RCHO
7: RCOR
8: RCOOH
9: RCOOR
10: ROR
11: RCCH
12:RCN

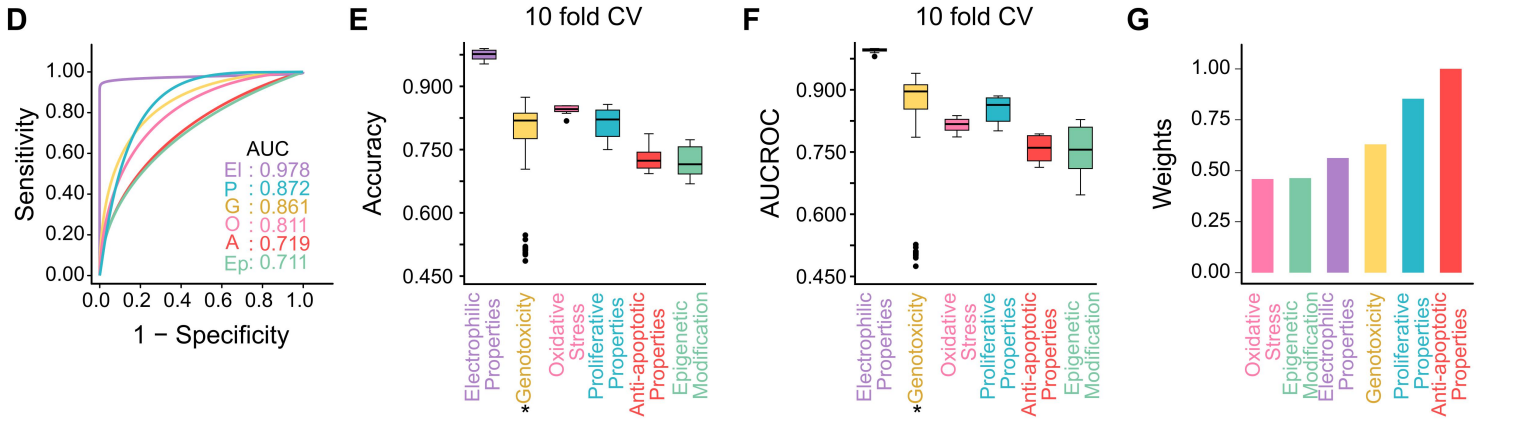

D
Sensitivity
1 − Specificity
AUC
El : 0.978
P : 0.872
G : 0.861
O : 0.811
A : 0.719
Ep: 0.711
E
10 fold CV
Accuracy
F
10 fold CV
AUCROC
Electrophilic Properties
*Genotoxicity
Oxidative Stress
Proliferative Properties
Anti-apoptotic Properties
Epigenetic Modification
G
Weights
Oxidative Stress
Epigenetic Modification
Electrophilic Properties
Genotoxicity
Proliferative Properties
Anti-apoptotic Properties

Figure 1

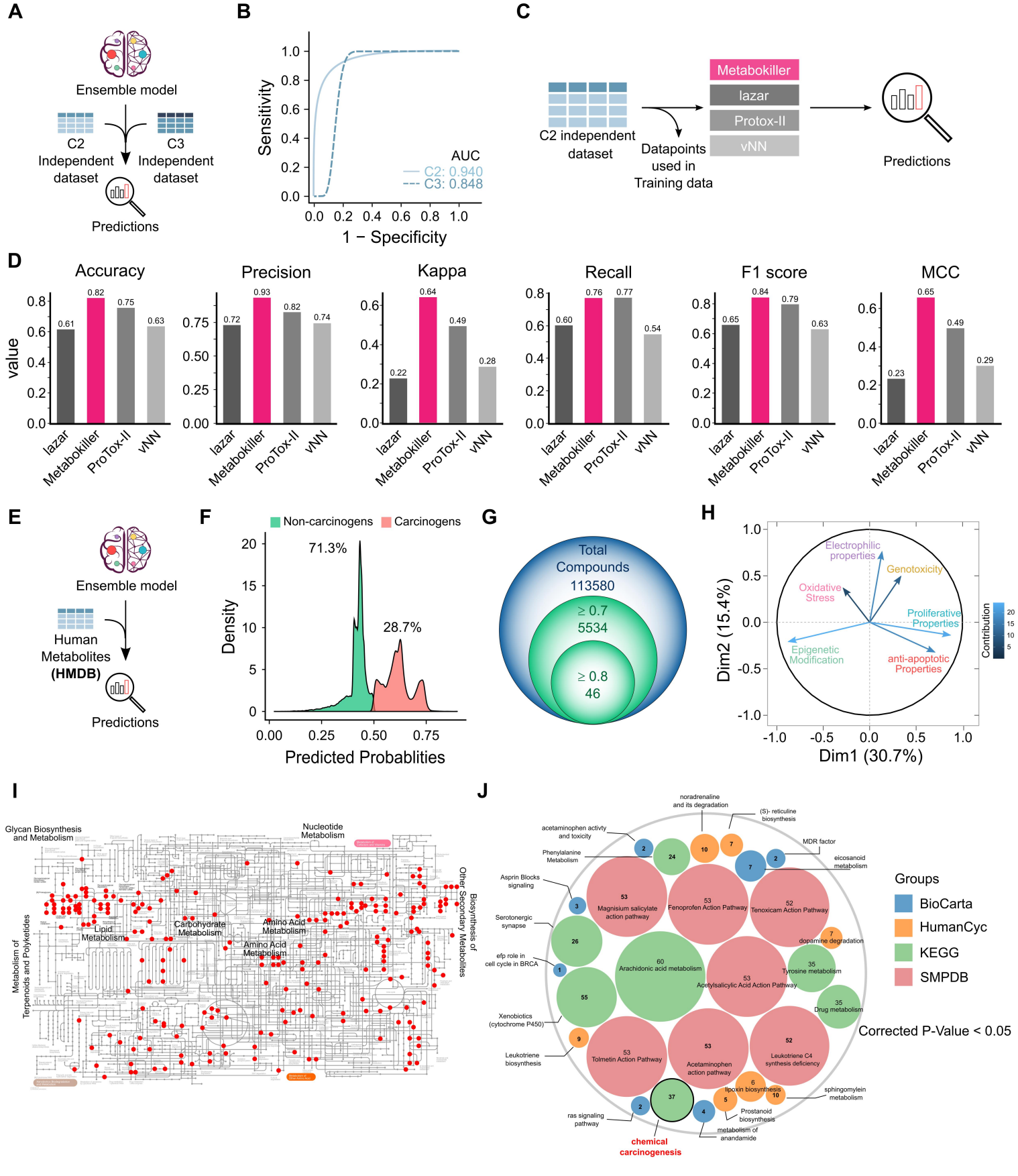

A
Ensemble model
C2
Independent
dataset
C3
Independent
dataset
Predictions
B
Sensitivity
1 − Specificity
AUC
C2: 0.940
C3: 0.848
C
C2 independent
dataset
Datapoints
used in
Training data
Metabokiller
lazar
Protox-II
vNN
Predictions
D
Accuracy
Precision
Kappa
Recall
F1 score
MCC
value
lazar
Metabokiller
ProTox-II
vNN
0.61
0.82
0.75
0.63
0.72
0.93
0.82
0.74
0.22
0.64
0.49
0.28
0.60
0.76
0.77
0.54
0.65
0.84
0.79
0.63
0.23
0.65
0.49
0.29
E
Ensemble model
Human
Metabolites
(HMDB)
Predictions
F
Non-carcinogens
Carcinogens
71.3%
28.7%
Density
Predicted Probablities
G
Total
Compounds
113580
≥ 0.7
5534
≥ 0.8
46
H
Electrophilic
properties
Genotoxicity
Oxidative
Stress
Proliferative
Properties
Epigenetic
Modification
anti-apoptotic
Properties
Contribution
Dim2 (15.4%)
Dim1 (30.7%)
I
Glycan Biosynthesis
and Metabolism
Nucleotide
Metabolism
Lipid
Metabolism
Carbohydrate
Metabolism
Amino Acid
Metabolism
Amino Acid
Metabolism
Metabolism of
Terpenoids and Polyketides
Biosynthesis of
Other Secondary Metabolites
J
noradrenaline
and its degradation
(S)- reticuline
biosynthesis
acetaminophen activty
and toxicity
MDR factor
Phenylalanine
Metabolism
eicosanoid
metabolism
Asprin Blocks
signaling
Serotonergic
synapse
efp role in
cell cycle in BRCA
Xenobiotics
(cytochrome P450)
Leukotriene
biosynthesis
ras signaling
pathway
chemical
carcinogenesis
metabolism of
anandamide
Prostanoid
biosynthesis
lipoxin biosynthesis
sphingomylein
metabolism
dopamine degradation
Magnisium salicylate
action pathway
Fenoprofen Action Pathway
Tenoxicam Action Pathway
Arachidonic acid metabolism
Acetylsalicylic Acid Action Pathway
Tyrosine metabolism
Drug metabolism
Tolmetin Action Pathway
Acetaminophen
action pathway
Leukotriene C4
synthesis deficiency
Groups
BioCarta
HumanCyc
KEGG
SMPDB
Corrected P-Value < 0.05

Figure 2

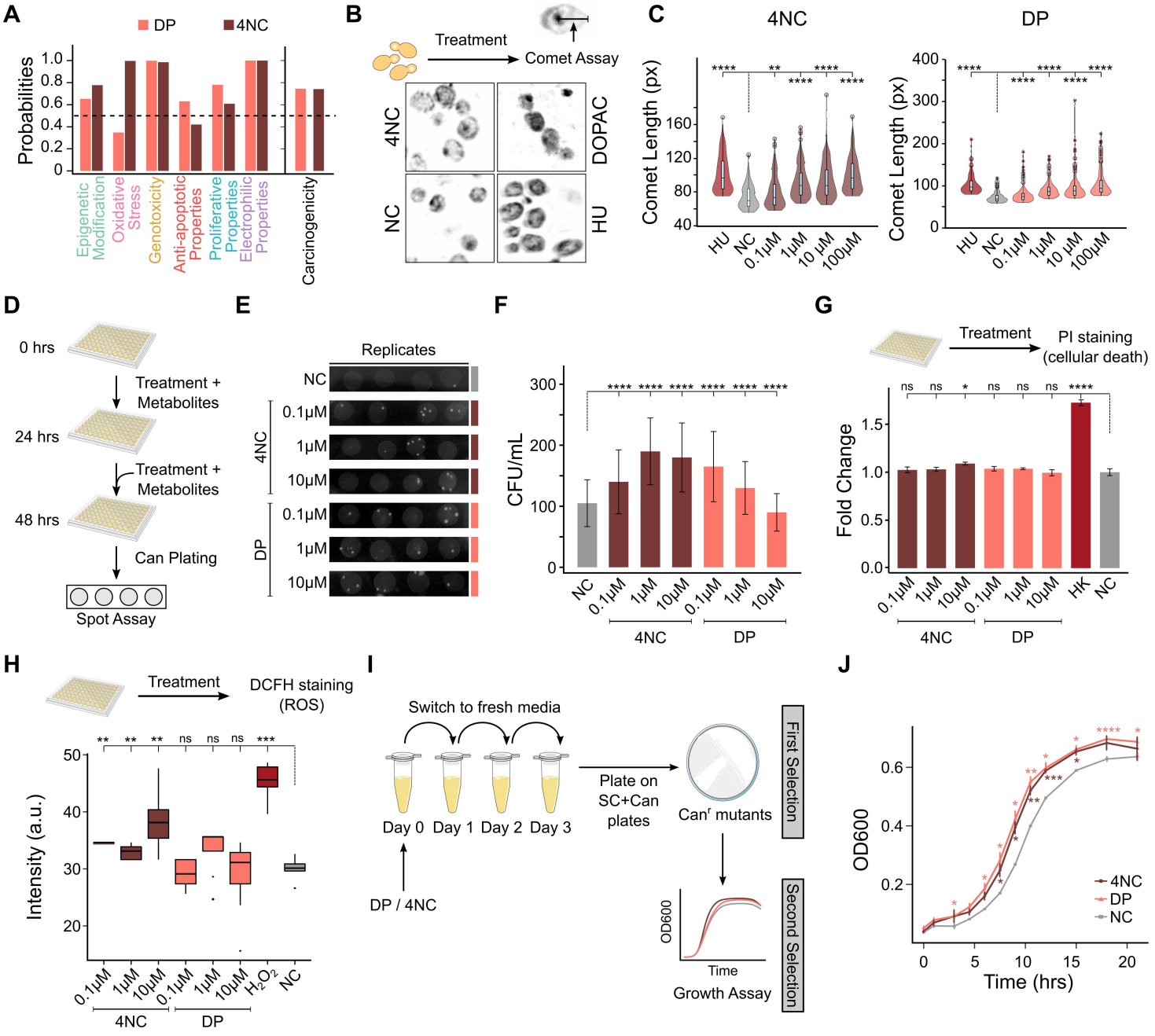

A
DP
4NC
Probabilities
Epigenetic Modification
Oxidative Stress
Genotoxicity
Anti-apoptotic Properties
Proliferative Properties
Electrophilic Properties
Carcinogenicity
B
Treatment
Comet Assay
4NC
NC
DOPAC
HU
C
4NC
DP
Comet Length (px)
HU
NC
0.1μM
1μM
10 μM
100μM
D
0 hrs
Treatment + Metabolites
24 hrs
Treatment + Metabolites
48 hrs
Can Plating
Spot Assay
E
Replicates
NC
4NC
0.1μM
1μM
10μM
DP
F
CFU/mL
G
Treatment
PI staining (cellular death)
Fold Change
HK
H
Treatment
DCFH staining (ROS)
Intensity (a.u.)
$H_2O_2$
I
Switch to fresh media
Day 0
Day 1
Day 2
Day 3
DP / 4NC
Plate on SC+Can plates
Can$^r$ mutants
First Selection
OD600
Time
Growth Assay
Second Selection
J
OD600
Time (hrs)

Figure 3

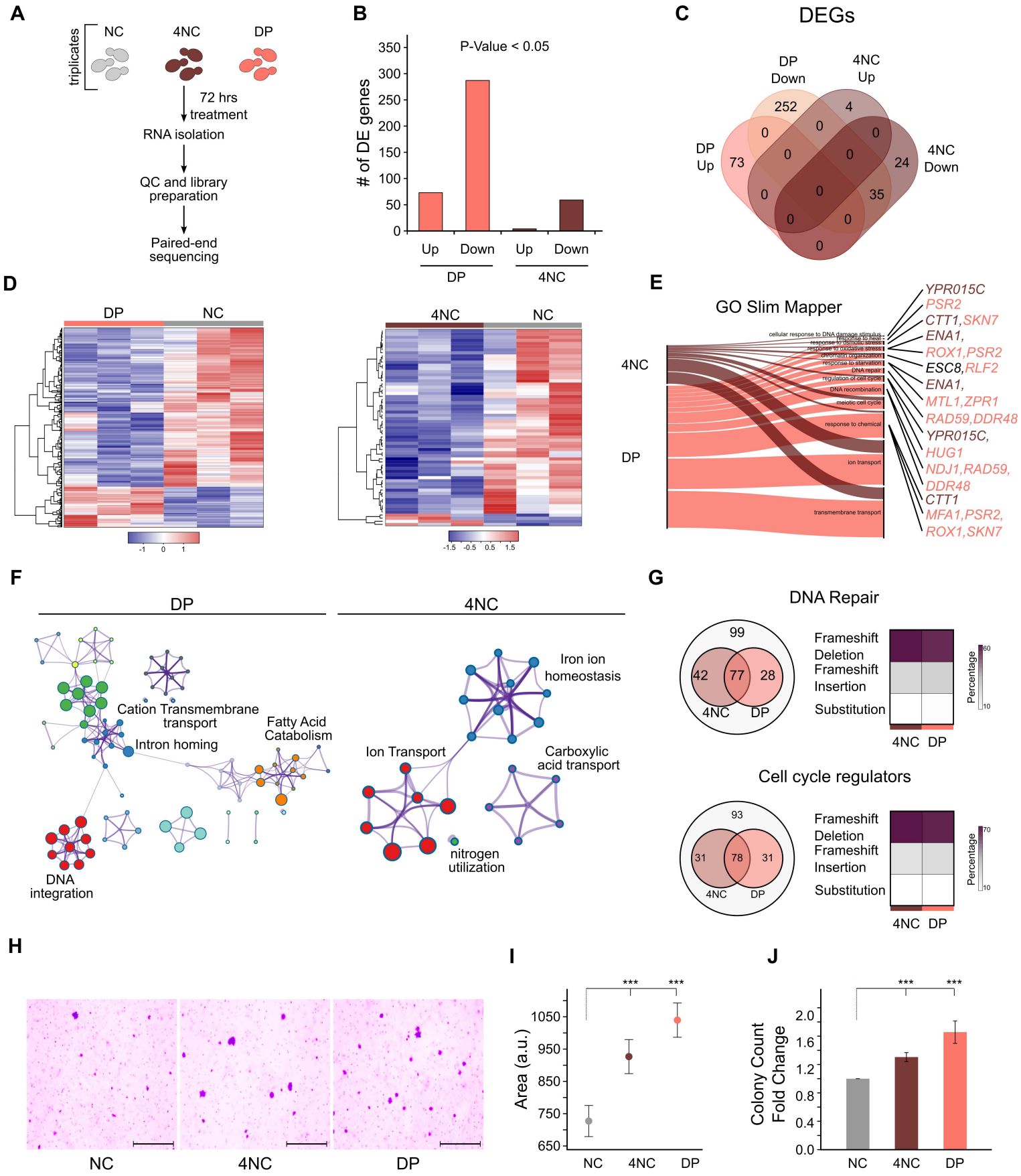

A
triplicates
NC
4NC
DP
72 hrs treatment
RNA isolation
QC and library preparation
Paired-end sequencing
B
P-Value < 0.05
# of DE genes
350
300
250
200
150
100
50
0
Up
Down
Up
Down
DP
4NC
C
DEGs
DP Down
4NC Up
DP Up
4NC Down
252
4
73
24
35
0
D
DP
NC
4NC
NC
-1
0
1
-1.5
-0.5
0.5
1.5
E
GO Slim Mapper
4NC
DP
cellular response to DNA damage stimulus
response to heat
response to chemical stress
response to oxidative stress
chromatin organization
response to starvation
DNA repair
regulation of cell cycle
DNA recombination
meiotic cell cycle
response to chemical
ion transport
transmembrane transport
YPR015C
PSR2
CTT1,SKN7
ENA1,
ROX1,PSR2
ESC8,RLF2
ENA1,
MTL1,ZPR1
RAD59,DDR48
YPR015C,
HUG1
NDJ1,RAD59,
DDR48
CTT1
MFA1,PSR2,
ROX1,SKN7
F
DP
4NC
Cation Transmembrane transport
Intron homing
Fatty Acid Catabolism
DNA integration
Iron ion homeostasis
Ion Transport
Carboxylic acid transport
nitrogen utilization
G
DNA Repair
99
42
77
28
4NC
DP
Frameshift Deletion
Frameshift Insertion
Substitution
Percentage
60
10
4NC
DP
Cell cycle regulators
93
31
78
31
4NC
DP
Frameshift Deletion
Frameshift Insertion
Substitution
Percentage
70
10
4NC
DP
H
NC
4NC
DP
I
***
***
Area (a.u.)
1050
950
850
750
650
NC
4NC
DP
J
***
***
Colony Count Fold Change
2.0
1.6
1.2
0.8
0.4
0.0
NC
4NC
DP

Figure 4

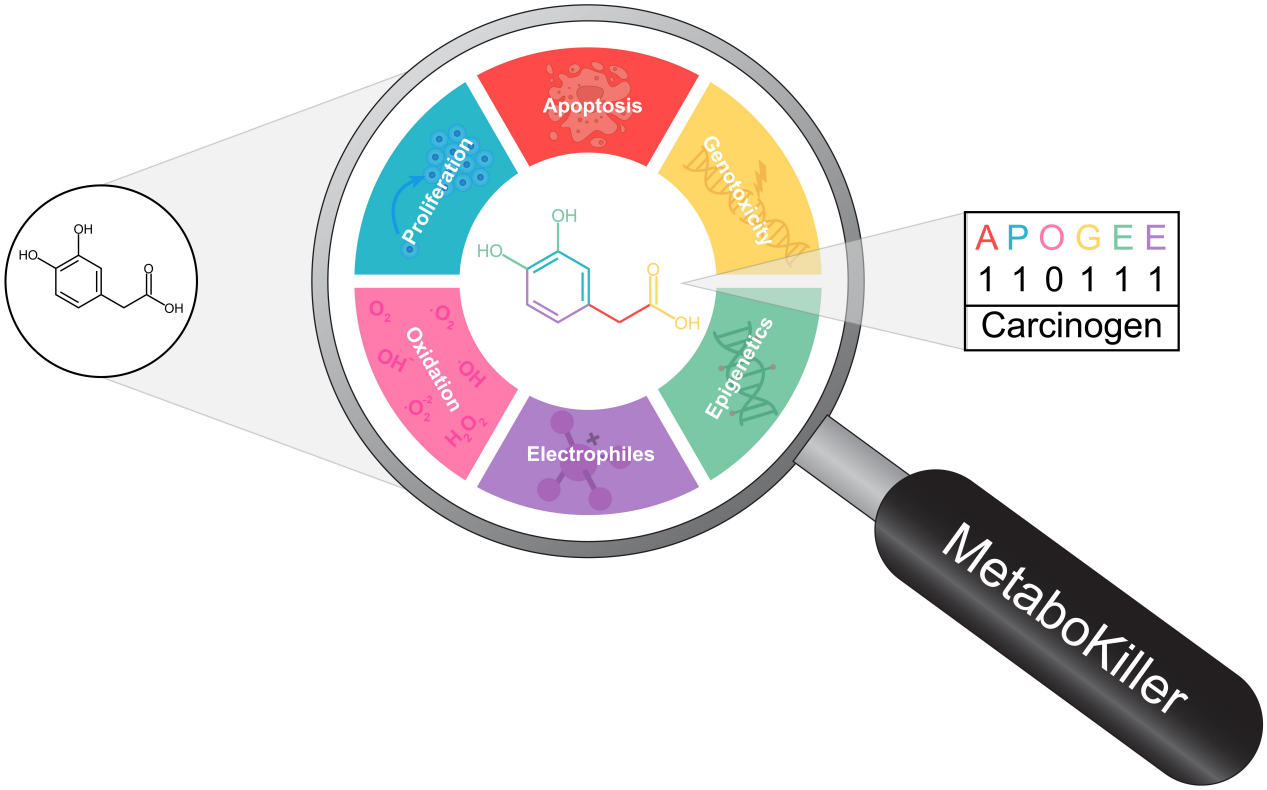

Apoptosis
Genotoxicity
Epigenetics
Electrophiles
Oxidation
Proliferation
OH
HO
OH
O
A P O G E E
1 1 0 1 1 1
Carcinogen
MetaboKiller

Figure 5

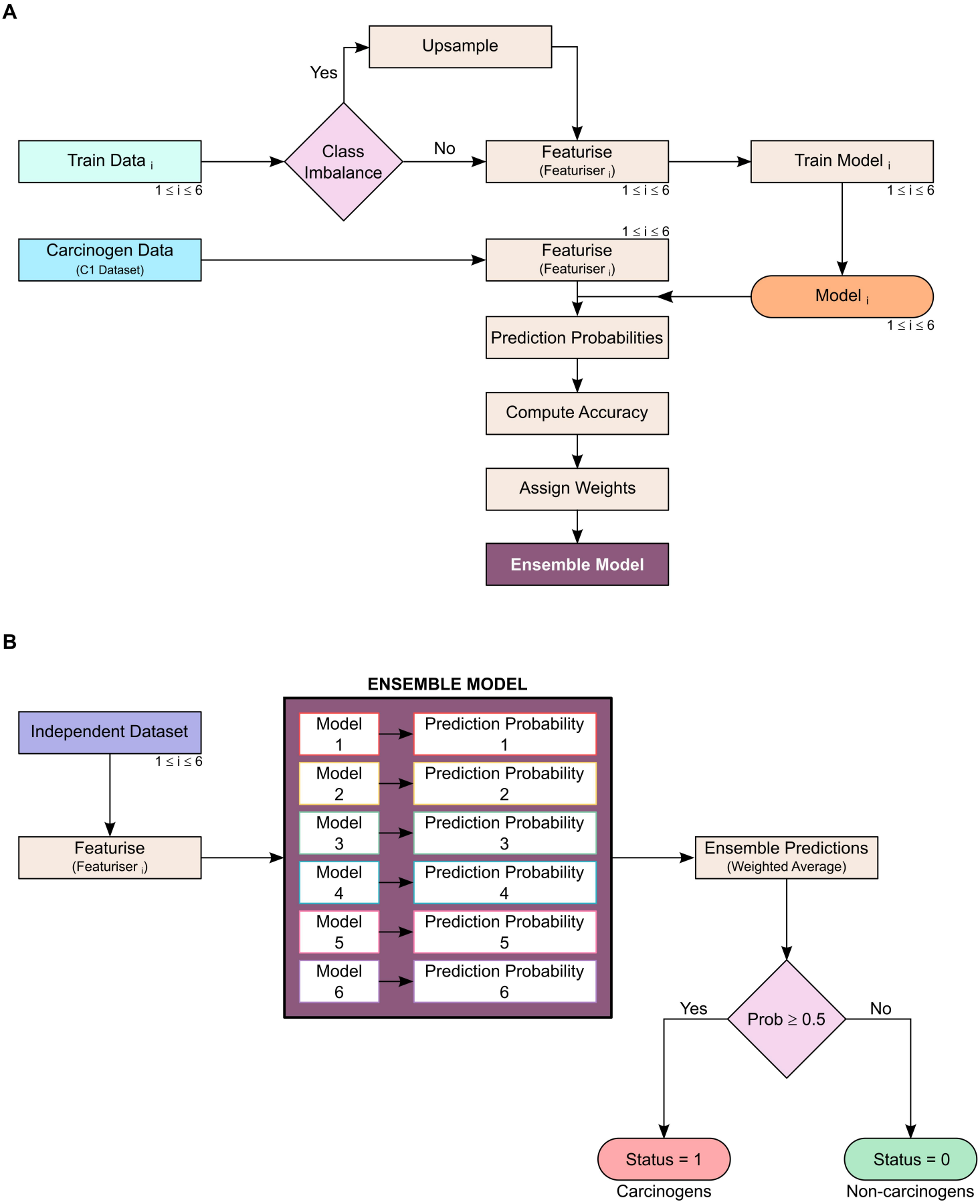
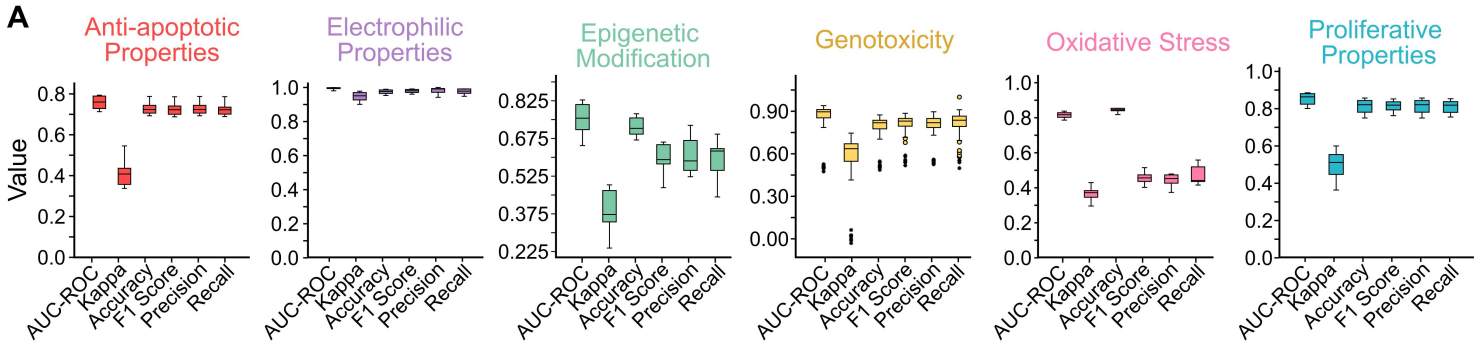

A
Upsample
Yes
Train Data i
1 ≤ i ≤ 6
Class Imbalance
No
Featurise
(Featuriser i)
1 ≤ i ≤ 6
Train Model i
1 ≤ i ≤ 6
Carcinogen Data
(C1 Dataset)
1 ≤ i ≤ 6
Featurise
(Featuriser i)
Model i
1 ≤ i ≤ 6
Prediction Probabilities
Compute Accuracy
Assign Weights
Ensemble Model
B
ENSEMBLE MODEL
Independent Dataset
1 ≤ i ≤ 6
Featurise
(Featuriser i)
Model 1
Prediction Probability 1
Model 2
Prediction Probability 2
Model 3
Prediction Probability 3
Model 4
Prediction Probability 4
Model 5
Prediction Probability 5
Model 6
Prediction Probability 6
Ensemble Predictions
(Weighted Average)
Yes
Prob ≥ 0.5
No
Status = 1
Carcinogens
Status = 0
Non-carcinogens

Figure S1

**A**




**B**

| Properties | Model | Feature Descriptors |
|---|---|---|
| Epigenetic Modification | Graph Convolution Network | MolGraphConv |
| Genotoxicity | GraphConvModel | ConvMolFeaturiser |
| Oxidative Stress | Attentive FP | MolGraphConv |
| Anti-apoptotic Properties | Random Forest | Mordred Features |
| Proliferative Properties | Extra Tree Classifier | Mordred Features |
| Electrophilic Properties | Attentive FP | MolGraphConv |

**C**

| Electrophilic Properties | | |
|---|---|---|
| Samples (1917) | Class 0 | 757 |
| | Class 1 | 1160 |
| Hyperparameters | num_layers | 2 |
| | graph_feat_size | 200 |
| | dropout | 0 |
| | batch_size | 32 |
| | learning_rate | 0.001 |

| Epigenetic Modification | | |
|---|---|---|
| Samples (753) | Class 0 | 493 |
| | Class 1 | 260 |
| Hyperparameters | graph_conv_layers | [64,64] |
| | predictor_hidden_feats | 64 |
| | predictor_dropout | 0.2 |
| | Learning_rate | 0.001 |
| | batch_size | 32 |

| Oxidative Stress | | |
|---|---|---|
| Samples (5548) | Class 0 | 4774 |
| | Class 1 | 774 |
| Hyperparameters | num_layers | 2 |
| | graph_feat_size | 200 |
| | dropout | 0.5 |
| | Learning_rate | 0.001 |
| | batch_size | 32 |

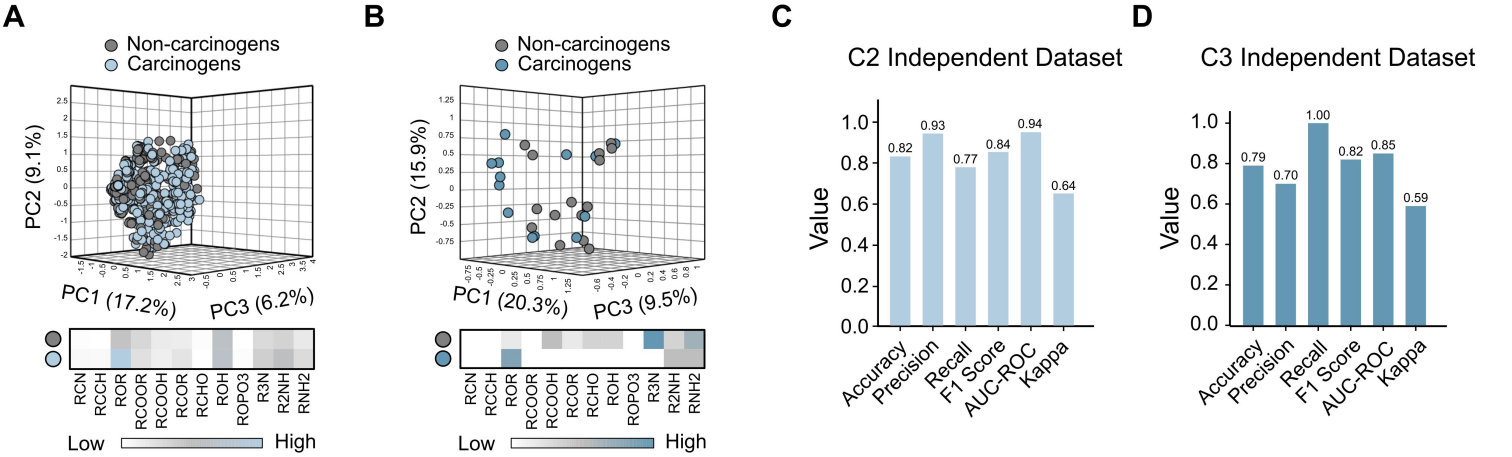
| Anti-apoptotic Properties | | |
|---|---|---|
| Samples (569) | Class 0 | 353 |
| | Class 1 | 216 |
| Hyperparameters | n_estimators | 100 |
| | min_samples_split | 2 |
| | min_samples_leaf | 1 |
| | max_features | sqrt |
| | max_depth | 100 |
| | bootstrap | False |

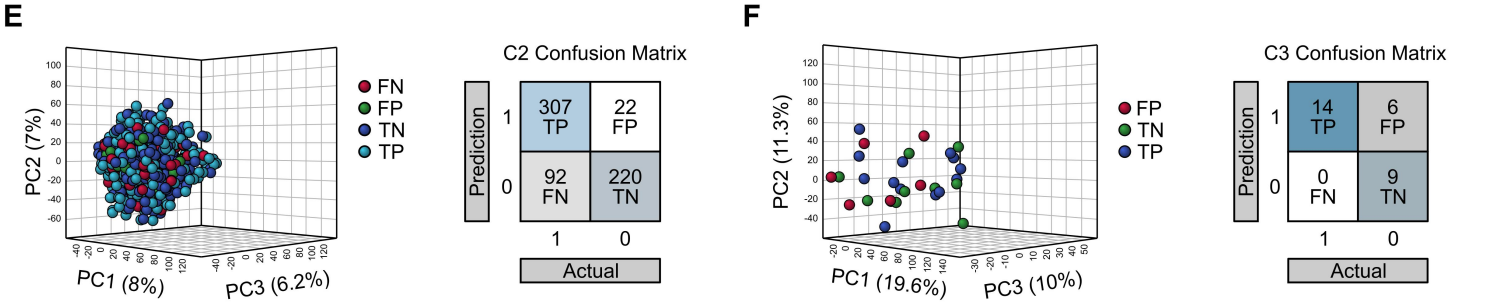
| Genotoxicity | | |
|---|---|---|
| Samples (6542) | Class 0 | 3026 |
| | Class 1 | 3516 |
| Hyperparameters | n_tasks | 1 |
| | batch_size | 32 |
| | learning_rate | 0.001 |

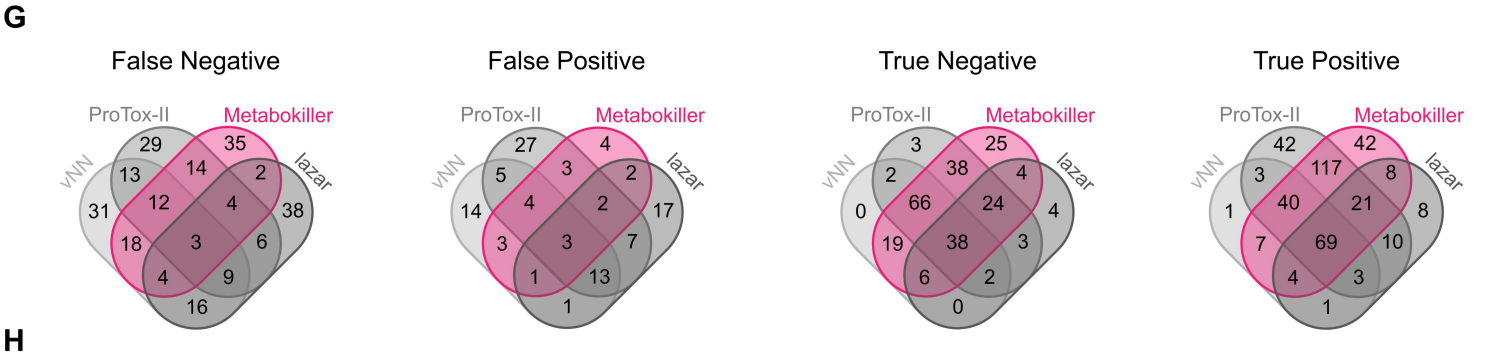
| Proliferative Properties | | |
|---|---|---|
| Samples (280) | Class 0 | 210 |
| | Class 1 | 70 |
| Hyperparameters | n_estimators | 90 |
| | max_depth | 15 |

Class 0 = anti- or no-activity
Class 1 = pro-activity

Figure S2




H

| Authors | Data Type | Features | Method | Accuracy |
|---|---|---|---|---|
| Fjodorova et al. | Carcinogen and Non-carcinogen | Molecular Descriptors | CP-ANN (Artificial Neural Network) | 68.3% |
| Zhang et al. | Carcinogen and Non-carcinogen | Molecular Descriptors, ECFP Fingerprints | Naïve Bayes* | 68% |
| Li et al. | Carcinogen and Non-carcinogen | MACCS Fingerprints | k-Nearest Neighbour (kNN)* | 80.46% |
| Zhang et al. (CarcinoPred-EL) | Carcinogen and Non-carcinogen | PaDEL Descriptors | Ensemble XGBoost* | 70.1% |
| Banerjee et al. (ProTox-II) | Carcinogen and Non-carcinogen | MACCS Fingerprints, Morgan Fingerprints | Random Forest | 83.3% |
| Maunz et al. (lazar) | Mutagen and Non-mutagen | BBRC Descriptors | Support Vector Machine (SVM)* | 74.6% |
| Schyman et al. (vNN) | Mutagen and Non-mutagen | ECFP Fingerprints | Variable Nearest Neighbor (vNN)* | 79% |

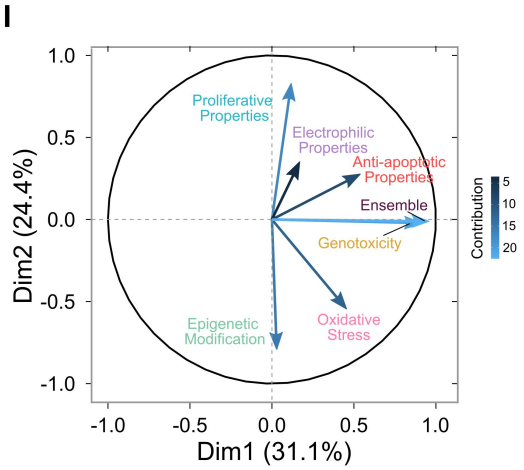



Figure S3

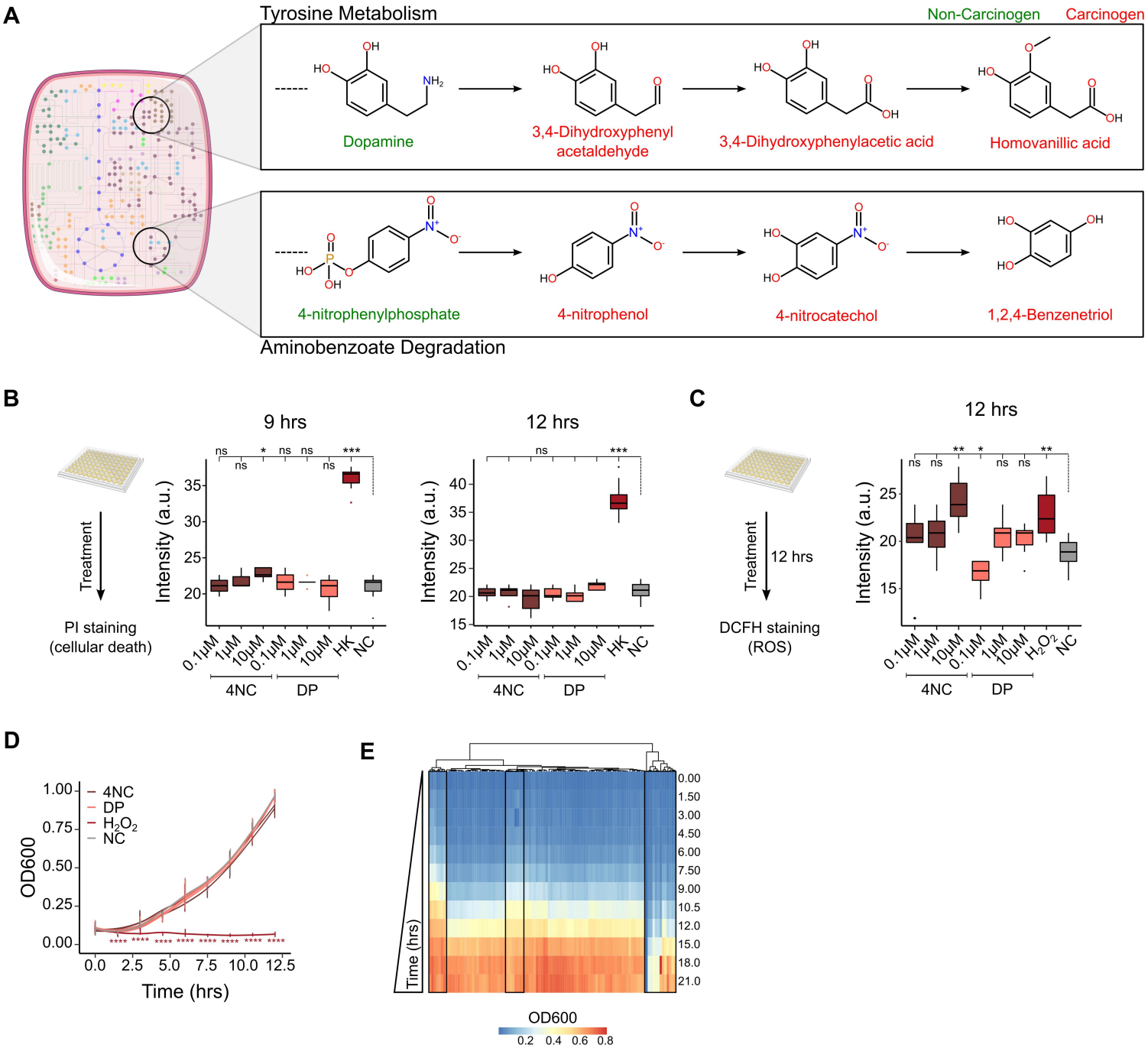

A
Tyrosine Metabolism
Non-Carcinogen
Carcinogen
Dopamine
3,4-Dihydroxyphenyl acetaldehyde
3,4-Dihydroxyphenylacetic acid
Homovanillic acid
4-nitrophenylphosphate
4-nitrophenol
4-nitrocatechol
1,2,4-Benzenetriol
Aminobenzoate Degradation
B
9 hrs
12 hrs
Treatment
PI staining (cellular death)
Intensity (a.u.)
0.1μM
1μM
10μM
HK
NC
4NC
DP
C
12 hrs
DCFH staining (ROS)
H2O2
D
4NC
DP
H2O2
NC
OD600
Time (hrs)
E
Time (hrs)
OD600

Figure S4

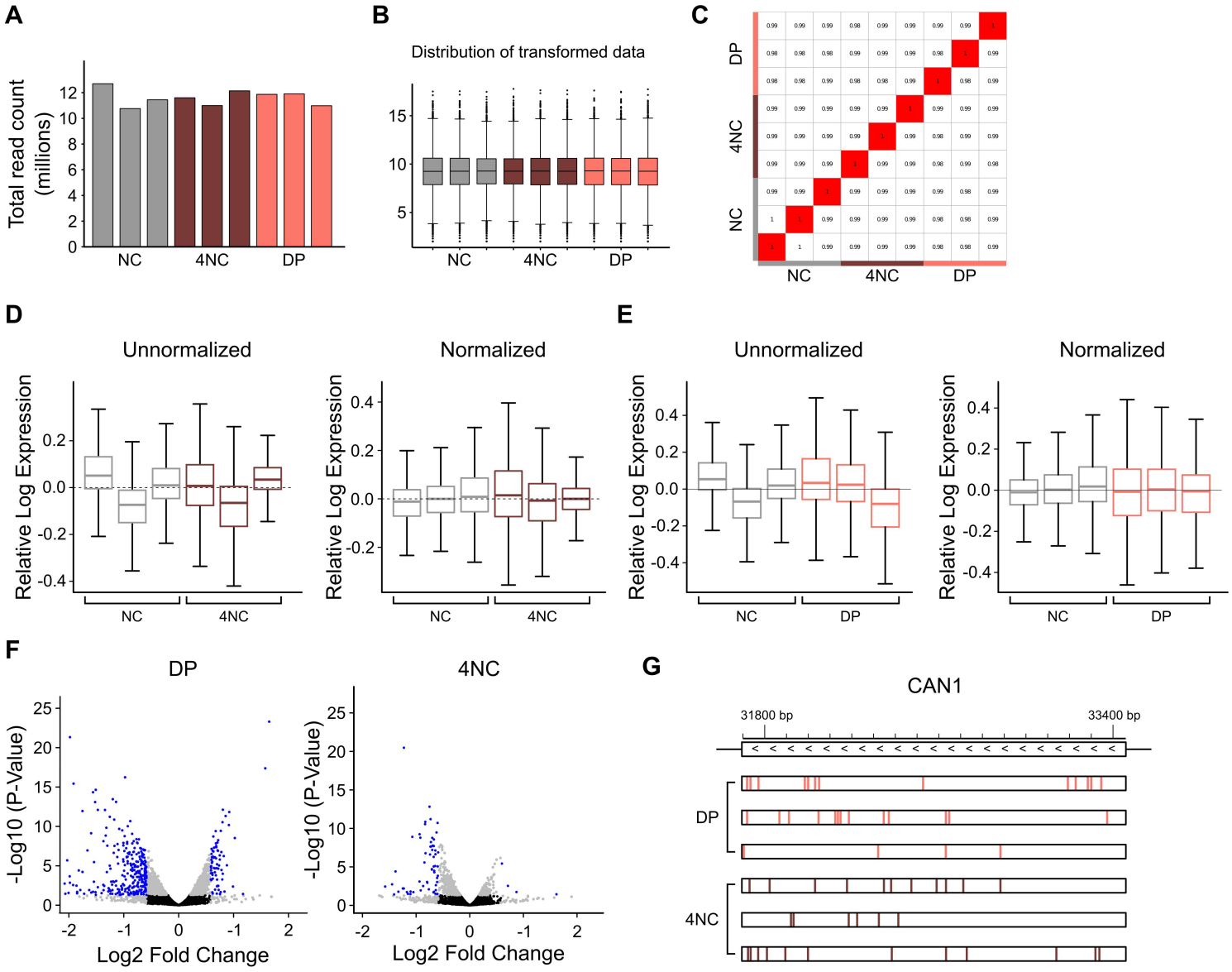

A
Total read count (millions)
NC
4NC
DP
B
Distribution of transformed data
NC
4NC
DP
C
DP
4NC
NC
NC
4NC
DP
D
Unnormalized
Normalized
Relative Log Expression
NC
4NC
E
Unnormalized
Normalized
Relative Log Expression
NC
DP
F
DP
4NC
-Log10 (P-Value)
Log2 Fold Change
G
CAN1
31800 bp
33400 bp
DP
4NC

Figure S5