Sparse Modeling of Genomic Landscape Identifies Pathogenic Processes and Therapeutic Targets in Metastatic Breast Cancer

Mengchen Pu[1, +, *], **Kai Tian**[1, +], **Weisheng Zheng**[1, +], **Xiaorong Li**[1, 2], **Keyue Fan**[1, 3], **Liang Zheng**[1], **Jielong Zhou**[1,*], **and Yingsheng Zhang**[1,*]

[1] StoneWise, AI, Ltd., Beijing, China

[2] Minzu University of China, Beijing, China

[3] Capital Normal University, Beijing, China

[*] Correspondence to: pumengchen@stonewise.cn (0000-0001-6282-2454) and zhangyingsheng@stonewise.cn (ORCID: 0000-0003-2520-3923)

[+] these authors contributed equally to this work

ABSTRACT

Breast cancer is a heterogeneous disease and ranks as one of the most lethal and frequently detected disease in the world. It poses significant challenges for precision therapy. To better decipher the patterns of heterogeneous nature in human genome and converge them into common functionals, mutational signatures are introduced to define the types of DNA damage, repair and replicative mechanisms that shape the genomic landscape of each cancer patient.

In this study, we developed a deep learning (DL) model, MetaWise 2.0, based on pruning technology that improved model generalization with deep sparsity. We applied it to patient samples from multiple sequencing studies, and identified statistically significant mutational signatures associated with metastatic progression using Shapley additive explanations (SHAP). We also employed gene cumulative contribution abundance analysis to link the mutational signatures with relevant genes, which could unearth the shared molecular mechanisms behind tumorigenesis and metastasis of each patient and lead to novel therapeutic target identification.

Our study illustrates that MetaWise 2.0 is an effective DL tool for discovering clinically meaningful mutational signatures in metastatic breast cancer (MBC) and relating them directly to relevant biological functions and gene targets. These findings could facilitate the development of novel therapeutic strategies and improve the clinical outcomes for individual patients.

Introduction

Cancer genome are prone to numerous mutations and rearrangements that manifest genomic instability and heterogeneity. These variants modulate the expression and function of genes that regulate cell growth, differentiation, survival and migration[1,2]. The predominant cause of cancer-related morbidity and mortality is the metastatic spread, in which cancer cells disseminate from their primary site to other parts of the body through blood or lymphatic vessels. Metastatic processes typically involve cellular stressors

and environmental shocks that elicit dramatic changes in the genome of cancer cells. These changes can bestow adaptive advantages to the cancer cells, such as enhanced invasiveness and therapeutic resistance. Therefore, elucidating the genomic characterization that underpins metastasis is crucial for devising effective strategies to prevent and treat cancers.

Breast cancer is the most common malignancy among women worldwide. For both primary and MBC, cumulative evidences point to the identification of the heterogeneous repertoire of disease-causing genes from various mutational processes[3,4]. However, due to the factors such as genomic background, lifestyle, tumor evolution and treatment pressure, the genomic alterations in metastatic breast cancer can differ drastically among patients. Thus, thorough genomic characterization of metastases for each patient will provide valuable insights, and is essential to understand the effects of systemic treatment on the tumor genome and improve the precision treatment of patients with metastatic breast cancers.

Genomic mutations can be characterized by mutational signatures defined as the proportion of mutations falling into mutational processes defined by their nucleotide context. Somatic mutations in whole genomes have empowered the detection of multiple mutational processes active in tumorigenesis, and such processes manifested in the tumor genome during tumor progression and treatment. Several studies have investigated the genomic landscape of metastatic tumors varieties[5-8]. Deep learning methods have been increasingly applied in the field of cancer research, particularly in studying the cancer metastasis progress. AI-based prediction models have been developed based on clinical data, such as medical images, gene expression profiles, *etc*[9-14]. At the same time, deep learning architectures are also being developed for the prediction of metastasis, including multilayer perceptron, convolutional neural networks, autoencoders, *etc*[15-18]. These models aim to solve a binary classification problem by classifying samples as either metastatic or non-metastatic. Several studies have shown the importance of mutational signatures in identifying cancer-associated genes and demonstrated some promising results on how these models guide precision medicine approaches[17,19]. However, major challenges remain to be addressed, including limited data accessibility, incorporation of multi-omics data, especially with extensive phenotypic annotations, limited generalizability across diverse patient cohorts, and the poor interpretability of DL models.

Over-parameterization, a common phenomenon in DL models, leads to high computational costs and impaired generalizability. Network pruning, the elimination of model components, has proven to be an effective technique that can optimize the efficiency of DL networks. This can help to reduce overfitting, improve model interpretability, and decrease computational requirements. In this study, we introduce an updated and extended version of MetaWise[17]. In comparison to our previous version, the new MetaWise 2.0 benefits from pruning technology, and offers model the generalization capability for data from different cohorts. Moreover, we integrated gene cumulative contribution abundance analysis with SHapley Additive exPlanations (SHAP) analysis to detect the significant correlations through association analysis of mutational signatures and key mutated functional genes and biological processes. These signatures were implicated in the growth of metastatic breast cancer cells, including those related to patient age, APOBEC enzymatic activity, DNA repair deficiency, *etc*. Several essential genes were identified to be the major biomarkers of MBC. Further enrichment analysis enabled the identification of various biological pathways involved in the development of MBC, such as the immune system processes, cell-cell communications, *etc*. The model also exhibited the capability to additionally stratify patients with MBC into more refined subgroups. This illustrates the potential utility of our updated approach, not only to discriminate the primary and metastatic cancers but also pinpoint the disease-associated molecular mechanisms for better therapeutic strategies of cancer treatment.

Results

Design of a sparse model for metastasis prediction

The diagram in Figure 1a illustrates the spread of MBC, originating from distinct primary sites and migrating toward several metastatic sites in our datasets. Apart from lymph nodes and breast as the

locoregional metastatic sites, the most frequently observed distant metastatic organs in our dataset were liver, bone, skin and lung. As shown in Figure 1b, the distribution of molecular subtypes in our dataset differed between primary and metastatic breast cancer patients. In primary patients, ER+/HER- and triple-negative breast cancer (TNBC) were the two most common subtypes, accounting for 36.87% and 36.41% of cases respectively. In metastatic patients, ER+/HER2- remained the most common subtype but with more predominant compared to primary cases, accounting for 59.77% of metastatic cases. The remaining distribution of metastatic patients consisted of 12.63%, 9.59% and 3.98% TNBC, ER+/HER2+ and ER-/HER2+ respectively.

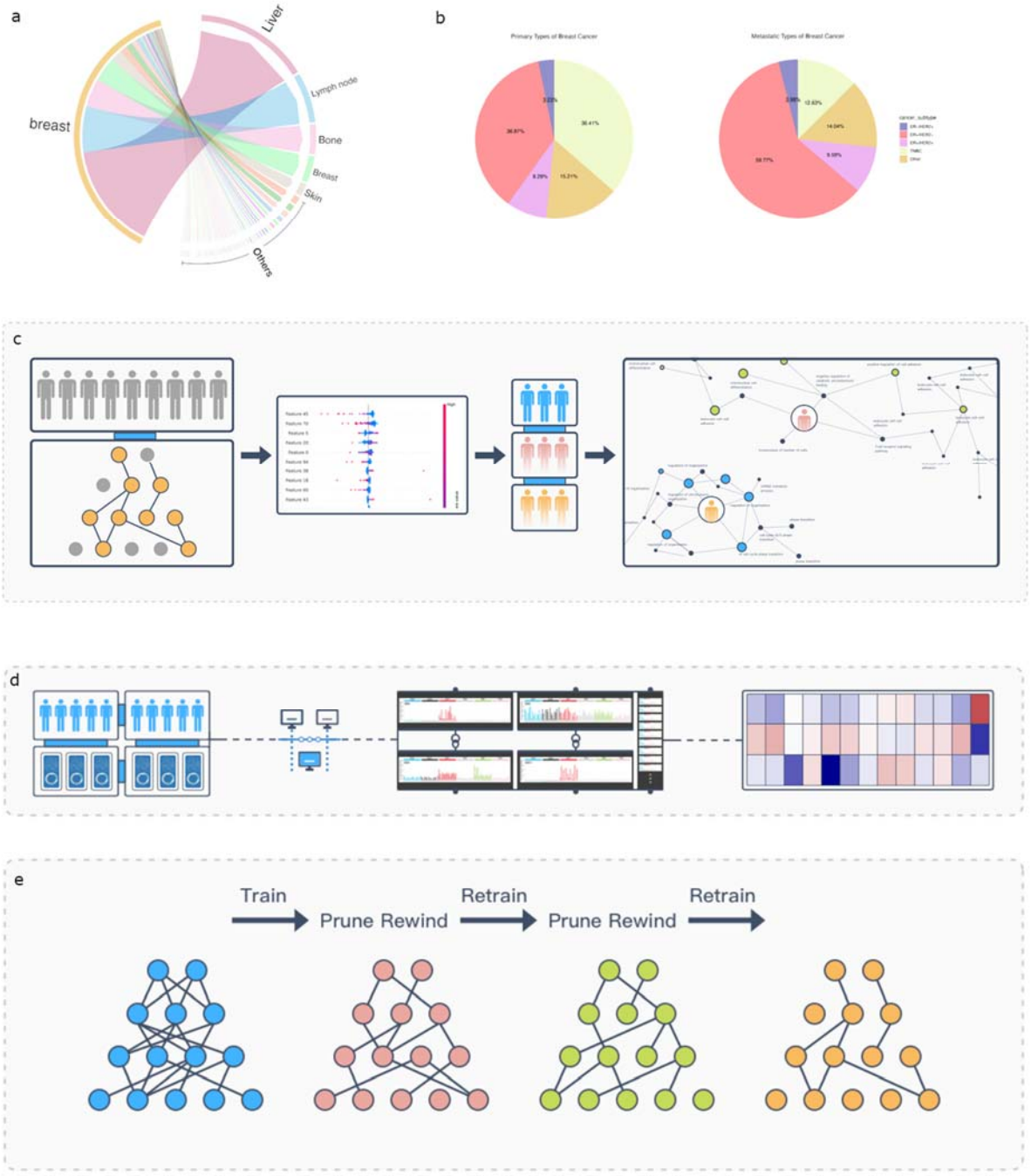


Figure 1. a. The Sankey diagram displays metastatic spreading directions from primary breast cancer toward several metastatic sites. Bandwidth is proportional to the number of metastatic tumor samples. Circle border thickness is proportional to the number of metastatic samples in that site. The color code representing the corresponding organ sites. b. The pie chart illustrates the subtypes distribution of primary and metastatic cancers within our dataset. The red, yellow, blue and grey color represents the ER+/HER2+, ER+/HER2-, ER-/HER2- and TNBC respectively. c. The workflow demonstrates our updated approach, MetaWise 2.0. d. The illustration of data pre-processing. e. The illustration of model pruning strategies.

As shown in Figure 1c, the workflow of our proposed model is illustrated, including a pruned deep learning predictor, an explanation module, and a sub-group gene enrichment analysis module. The predictor benefits from automated machine learning and model pruning techniques. To construct an optimal neural network architecture for predicting metastasis, we first utilized Bayesian optimization[20], an efficient tool for hyperparameter tuning that employs a probabilistic model to guide the search through the configuration space towards the globally optimal design. After conducting an automated architecture search to identify the ideal configurations for architectural aspects such as the number of hidden layers and nodes per layer, as well as fine-tuning the hyperparameters, we applied pruning to the model to optimize it in two stages: input feature pruning and global weight pruning. Input feature pruning seeks to decrease the dimensionality of the input data by eliminating any irrelevant or redundant features. Global weight pruning takes a broader approach by removing weights from layers across the full model architecture. We used the magnitude-based criterion to rank the weights according to their absolute values, then removed a fixed percentage of weights with the smallest magnitudes, starting from 10% and increasing by 10% until 90%. We fine-tuned the sparse model for several epochs after each pruning step as shown in Figure 1e.

Leveraging from Sparse model improves the performance of MetaWise-BC

In comparison to the model prior to the implementation of pruning techniques, it has been observed that the performance of the model post-pruning exhibits a marginal improvement with respect to the validation and internal test data. Furthermore, the model with pruned input features only or with pruned weights only showed slightly better performance in validation dataset than pruned input features and weights simultaneously, as shown in Table 1. This suggests that the pruning process may have facilitated the optimization of the model, enabling it to more effectively learn from feature correlation and even make slightly better predictions based on the training data with re-training after each pruning step.

Upon application to an external test data, it was observed that the three distinct pruning models exhibited substantial improvements in performance compared to their pre-pruning counterparts. Notably, there was a marked increase in recall, with an enhancement of 12%, while the F1 score demonstrated a 4.9% increase, and test accuracy and AUC both improved by 3.6%, as shown in Table 2. Additionally, other metrics also improved in varying degrees. These results convincingly demonstrate the power of pruning techniques to reduce model parameters, and help to mitigate overfitting. Also, the incorporation of pruning allows the model to better capture the key feature correlations, generalize to new data cohorts and accurately predict a patient's potential risk of metastasis.

pruning	Val acc	Test acc	Test recall	Test precision	F1 score	Test AUC	Test AUPR
No	91.5%	86.9%	90.2%	94.3%	92.2%	78.4%	96.5%

Weight pruning	90.6%	87.9%	90.2%	95.4%	92.7%	81.8%	97%
Input pruning	91.5%	88.8%	92.4%	94.4%	93.4%	79.5%	96.7%
Weight + input pruning	91.5%	86%	89.1%	94.3%	91.6%	78.0%	96.4%

Table 1. The average training and validation performance of the model across five folds of cross-validation.

pruning	Test acc	Test recall	Test precision	F1 score	Test AUC	Test AUPR
No	82.6	77.6	86.4	81.7	82.6	87.6
Weight	85.5	86.9	84.7	85.7	85.5	89.1
Input	83.8	82.2	84.9	83.5	83.8	88
Weight + input	85.5	86	85.2	85.6	85.5	89.1

Table 2. The average performance on external data

SHAP analysis interprets the model

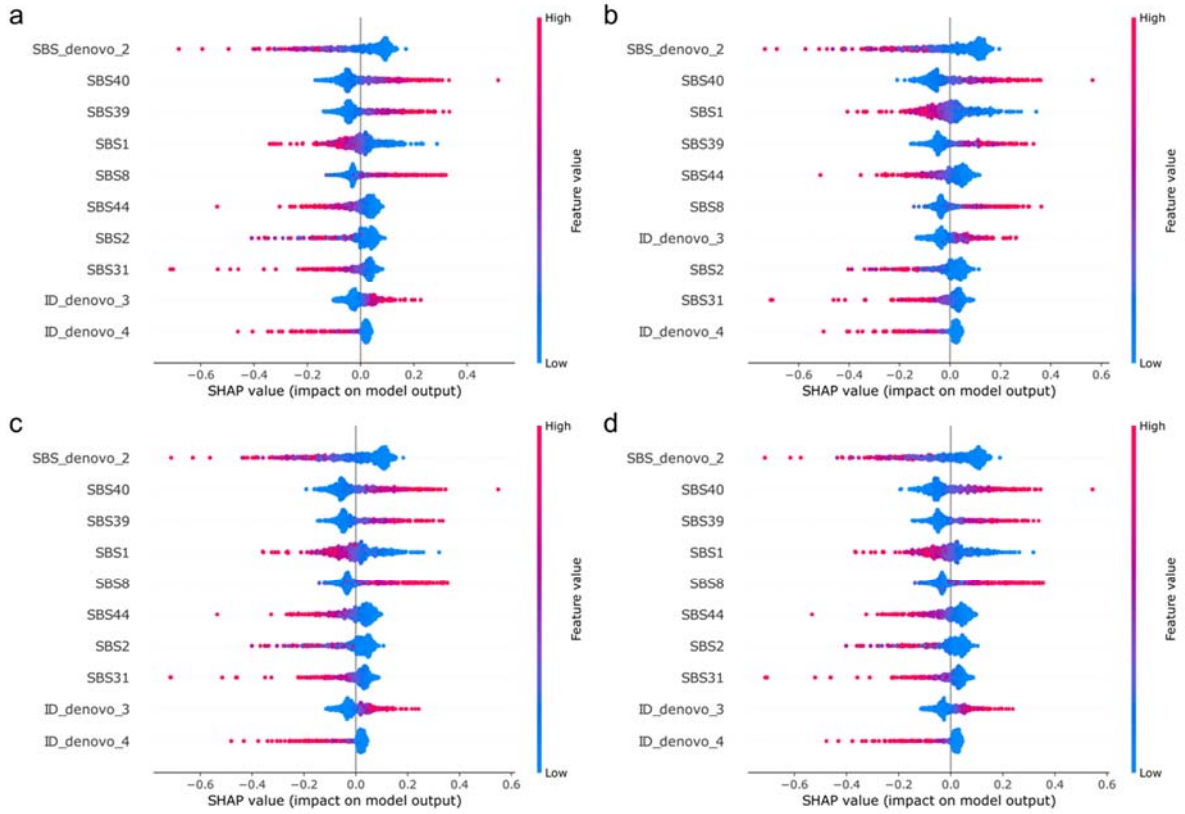
To investigate the role of mutational signatures in breast cancer metastasis, we performed SHAP analysis on our model and identified certain signatures, such as SBS40, SBS1, SBS39, SBS8, SBS44, SBS2, SBS31 and three de novo signatures SBS_denovo_2, ID_denovo_3 and ID_denovo_4 which are involved in the development of metastatic breast cancers, as shown in Supplementary Figure 1. In the pruned and unpruned models, SHAP analysis revealed nearly the identical ten most impactful signatures. The key factors were consistent across both models, only with varied relative importance. The clock-like signature SBS1 is attributed to endogenous deamination of 5-methylcytosine to thymine[21] which is related to age, as well as SBS40, which has also been shown to correlate with patient age in different types of human cancer[22]. The mutational signature SBS8 is common in most cancers, but its etiology is controversial. Recent evidence suggests that the SBS8 signature is due to DNA damage caused by late replication errors[23], similar with defective DNA mismatch repair related signature SBS44. The uncharacterized signature SBS39 was significantly enriched in the basal subtype compared with three other breast cancer subtypes defined by PAM50 (Her2, Luminal A, Luminal B) [24]. The SBS2 is associated with activity of the AID/APOBEC family of cytidine deaminases on the basis of similarities in the sequence context of cytosine mutations caused by APOBEC enzymes. The SBS31 is attributed to chemotherapy treatment with platinum drugs.

In order to examine the consistency between pruning strategies and SHAP analysis, we conducted a very radical pruning experiment by trimming 50% of the input features. The results showed that the eliminated features were consistently ranked lower in SHAP analysis with a Shapley value of 0.0 as shown in Supplementary Table 1. This suggests that the pruned input features had minimal impact on prediction outcomes, as determined by the interpretable analysis. Our findings indicate a high degree of consistency between the pruning strategy and SHAP analysis results, providing valuable insights into the effectiveness of these methods in improving model performance.

Mutation signature	Shapley value
SBS6	0.0
SBS7a	0.0
SBS7b	0.0
SBS12	0.0
SBS16	0.0
SBS17a	0.0
SBS19	0.0
SBS21	0.0
SBS22	0.0
SBS25	0.0
SBS26	0.0
SBS28	0.0
SBS30	0.0
SBS33	0.0
SBS35	0.0
SBS36	0.0
SBS38	0.0
SBS41	0.0
SBS88	0.0
SBS92	0.0
SBS93	0.0
SBS_denovo_5	0.0
SBS_denovo_6	0.0

SBS_denovo_8	0.0
SBS_denovo_9	0.0
SBS_denovo_10	0.0
SBS_denovo_11	0.0
SBS_denovo_12	0.0
SBS_denovo_13	0.0
SBS_denovo_14	0.0
SBS_denovo_15	0.0
SBS_denovo_16	0.0
SBS_denovo_20	0.0
SBS_denovo_21	0.0
SBS_denovo_22	0.0
SBS_denovo_23	0.0
SBS_denovo_24	0.0
SBS_denovo_25	0.0
SBS_denovo_26	0.0
SBS_denovo_27	0.0
DBS4	0.0
DBS5	0.0
DBS_denovo_5	0.0
DBS_denovo_7	0.0
ID3	0.0
ID_denovo_2	0.0
ID_denovo_5	0.0
ID_denovo_6	0.0
ID_denovo_7	0.0

Supplementary Table 1. The Shapley value of removed features by pruning.



Supplementary Figure 1. SHAP result of a. original model, b. Global weights pruned model, c. Input feature pruned model and, d. Global weights pruned and input pruned model.

To further validate the mutational signatures on model performance and results of the SHAP analysis, we conducted ablation studies. By systematically removing input mutation signatures with high Shapley value, we were able to assess the impact of these mutation signatures on model performance and understand the relationship between individual input mutation signatures and their contribution to prediction outcomes. The results of these ablation studies on top five features with large Shapley values are presented in Table 3. Our ablation studies support the effectiveness of SHAP analysis in identifying important mutational signatures and guiding the development of more accurate and interpretable models. After removing the top input features with the largest Shapley values, the models performance dropped significantly by approximately 5-10% on test accuracy. This suggests that the top features were critical to the model′s ability to predict. The sharp drop in performance after their removal highlights the importance of these features in the model and underscores the need for careful feature selection when developing predictive models.

Removed mutation signatures	Original	Weights	Input features	Weights and input features
None	**0.831**	**0.841**	**0.827**	**0.850**

SBS_denovo_2	0.803	0.822	0.817	0.822
SBS40	0.803	0.827	0.794	0.813
SBS1	0.757	0.785	0.771	0.752
SBS39	0.775	0.813	0.813	0.785
SBS8	0.738	0.761	0.752	0.742

Table 3. The model accuracy results from ablating the top features.

Cumulative contribution abundance analysis and GO enrichment analysis reveals potentially relevant features

We conducted a comprehensive analysis to investigate the molecular mechanism behind mutation signatures by employing the RNMF method[25] and gene enrichment analysis. A key component of this approach is the Cumulative Contribution Abundance (CCA) model, which effectively elucidates the associations between mutational signatures and genes. By calculating the cumulative contribution of each gene to each mutation signature, we were able to identify the genes that exerted the greatest influence on each mutation signature. This information was then applied to determine the genes that contributed most significantly to the mutation signature with the greatest impact for each sample. We subsequently grouped a subset of patients exhibiting similar characteristics of the most influential mutation signatures and obtained the most contributing genes corresponding to each patient within this subset. The resulting gene set was undertaken to gene enrichment analysis to further elucidate the underlying biological mechanisms.

Firstly, comparative analyses of the predicted genes enriched in primary and metastatic breast cancer samples were performed using KEGG, Reactome, and Gene Ontology databases (figure 2a). During the past decades, the molecular principles of metastasis remain an enigma even with the acceleration of multi-omics research[26,27]. Genetic immune escape (GIE)[28], microenvironment-derived epithelial to mesenchymal transition (EMT), cell motility[29,30], breast cancer stem cells' escape and sub localization are well characterized in the metastatic processes. In our study, the genes enriched in primary breast tumors are most correlated with cell growth and proliferation, cell homeostasis, and metabolism. Mutations of these genes can promote rapid cell proliferation, inhibit apoptosis, and ultimately lead to tumor formation. Nevertheless, the gene enriched in MBC are most correlated with immune system processes, cell communication, cell death *etc*. This implies these genes may participate in tumor metastasis and late-stage progression. The complement of these two sets showed significant differences in their corresponding biological functions. In the primary tumor enriched gene set, PI3K/AKT signaling, RAS-MAPK signaling, and ERBB signaling pathways were indicators of cell survival and proliferation; p53 signaling pathway, G1/S transition, and apoptotic response will affect cell death and cell cycle; carbon and lipids metabolism abnormality were associated with microenvironments, such as hypoxia and oxidative stress[31]. In comparison, four distinct pathways, including CCR3 pathway in inflammatory responses, PKC pathway, G protein signaling pathways and integration of energy metabolism, were specific to the metastasis enriched gene set. These pathways are associated with immune system processes, cell transformation and invasion, *etc*[32].

To further decipher the functions of these two distinct gene sets, we integrated them into a protein-protein physical interaction network (PPIN) and filtered the interactions with a confidence score of less than 2. The resulting genes were divided into six groups. As shown in Figure 2b, among these groups, five groups contained both primary and metastatic tumor genes, while the other group contained only primary tumor genes. The first group refers to cellular response to stimuli, especially immune response. The primary genes were mainly involved in RTK signaling and transcription regulation. In contrast, the metastasis genes were mainly involved in cellular surface communication and collagen recognition. The second group refers to transcription regulation. The primary tumor gene set was enriched in P53 signaling and hormone stimuli, while the metastasis genes were enriched in DNA repair and interleukin signaling. The other three groups of metastasis gene set were mainly enriched in Wnt signaling, RAP1 signaling, and calcium signaling pathway. The metastasis-specific genes were re-analyzed to characterize their functions in cellular processes. Nine genes, including growth factors, tyrosine kinases, GTPases, transcription factors and collagens, were involved in oncogenesis pathway (Figure 2c). For example, a PDGF gradient will drive cells to migrate towards the high concentration edge[33]; the extracellular matrix could be remodeled by different collagen types and concentrations to create a microenvironment supporting metastatic dissemination[34].

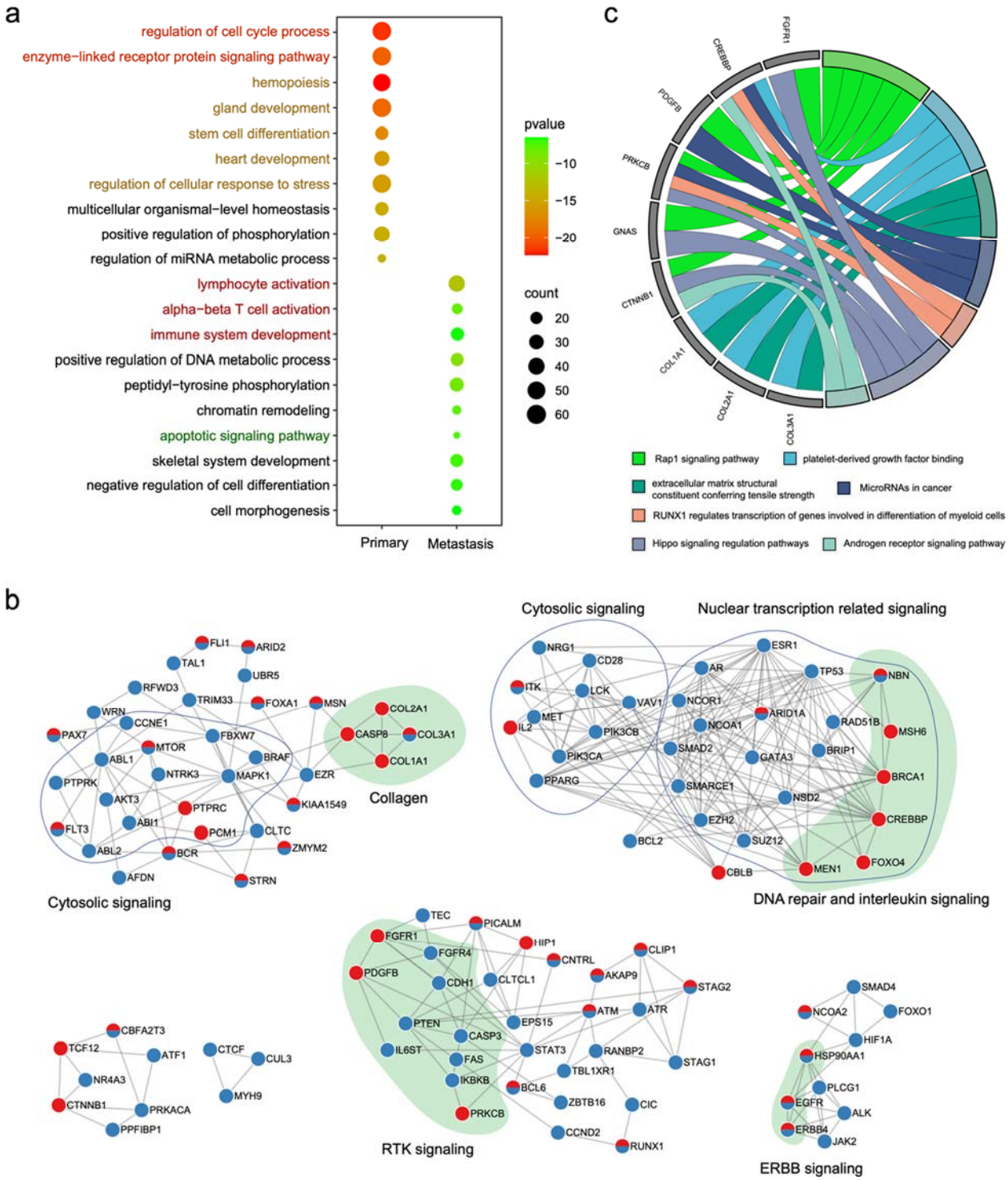


Figure 2. Visualizations of meta-analysis results of primary and metastasis gene sets. a. The gene ontology analysis of primary and metastasis enriched gene sets. b. The PPIN analysis of primary and metastasis enriched gene sets. c. The cellular processes analysis of nine metastasis specific genes.

To validate the potential of our model to effectively cluster patients, we analyzed the distinct groups within the metastatic patients. Eight clusters, including ID signature and SBS signature, were compared to

elucidate their functional divergence. 3/56 genes were assigned to two clusters, and some genes in different clusters play the same function (Figure 3a). The top 20 GO terms showed significant differences between different clusters, e.g., SBS1 is mainly enriched in miRNA transcription, while SBS2 is involved in protein-containing complex localization. Genes enriched in different clusters function in different biological processes (Figure 3b).

To better understand the mechanisms that differentiate the subgroups, we selected two distinct groups with the different highest impact mutation signatures: ID2 and SBS1. Genetic alterations analysis and PPIN analysis were performed for both sub-groups, with the results presented in Figure 3. Our findings revealed that the gene sets of these two patient subgroups were enriched in entirely distinct molecular functions. The SBS1 signature indicates a C or T substitution and is believed to result from different forms of DNA damage[35]. Ten genes were enriched in our prediction category. FOXO4, PAX3, PLAG1 and NFIB are transcription factors; PDGFB is an RTK-related signaling protein; GNAS is a G-protein downstream of GPCR; VTI1A and SNX29 are related to protein sorting; and MSH6 is a DNA-binding protein which facilitates DNA mismatch repair. The functional process of FOXO4, PAX3, GNAS and PDGFB would be linked to the AKT pathway (Figure 3c). PAX5, a paralog of PAX3, has been validated to induce the gene expression of E-cadherin[36] and MiR-215[37] to inhibit the expression of FAK[38], thereby suppressing breast cancer cell migration and invasion. A study of 263 breast cancer patients on cBioportal showed that 20 single missense mutations of PAX3 were related to metastatic breast cancer. Similarly, FOXO1 silencing in hepatocellular carcinoma causes ZEB2 expression and the EMT process[39], but high expression of FOXO1 and FOXO3 upregulates matrix metalloproteinase (MMP) expression and enhances cancer cell metastasis[40,41]. Missense mutations of FOXO1 or FOXO4 were also correlated with breast cancer metastasis according to genetic alterations analysis based on cBioportal database (Figure 3e).

ID1 and ID2 were the result of replication slippage with the most happening of A or T indels at long poly(dA:dT) tracts[42-44]. The predicted ID2 cluster including FLT3 and HSP90AA1 have functions in protein kinase activity and CTP binding. HSP90 functioned as protein chaperone might have its dual character in breast cancer oncogenesis: decreased HSP90 has documented to proceed invasion and metastasis, whilst increased HSP90 enhances cell proliferation[45]. Intriguingly, HSP90AA1 was confirmed to be secreted extracellularly, and could activate EMT and migration[46,47] (Figure 3d). These results indicate that our approach can successfully group patients in a way that reveals subsets enriched for distinct biological processes.

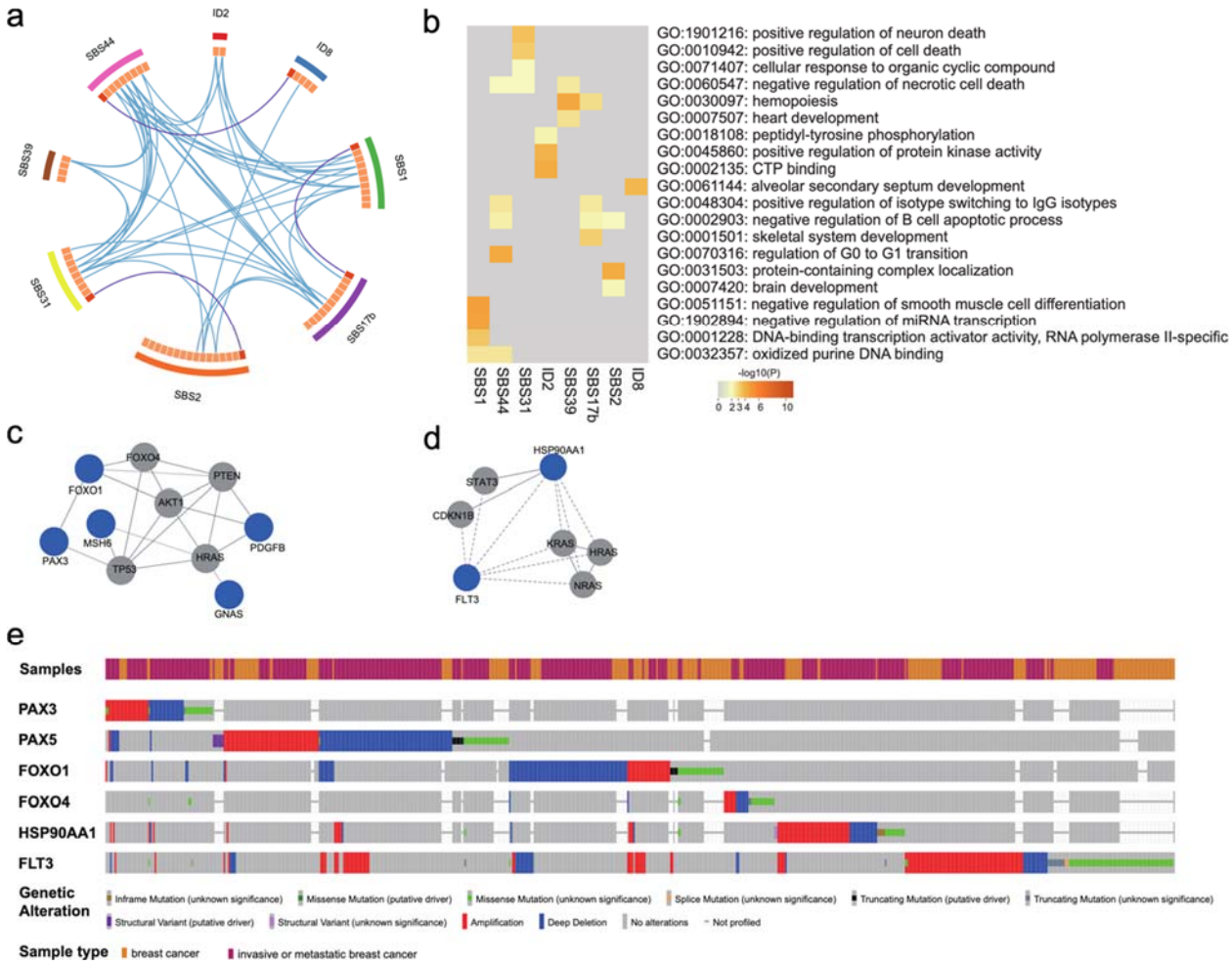


Figure 3. Visualizations of meta-analysis results based on multiple gene lists. A. predicted genes clustered well of sub-groups of metastatic patients. B. Heatmap showing the top enrichment clusters, one row per cluster, using a discrete color scale to represent statistical significance. Gray color indicates a lack of significance. C. The PPIN network of SBS1, the solid line represents the physical interactions which indicates that the proteins are part of a physical complex. The blue node represents the associated genes from mutational signatures. D. the PPIN network of ID1, the dot line represents the functional protein associations. E. Genetic alterations analysis of breast cancer with FOXO1, FOXO4, HSP90AA1 and FLT3.

Discussion

Over-parameterization is a common property in deep learning models, leading to increased computational costs and reduced generalization. As a remedy, network pruning has proven to be an effective technique to improve the efficiency of DL networks in situations where generalization is a concern and the computational cost is limited[48]. It is demonstrated in our work that pruning can be applied to the input layer of a neural network by removing input features with little to no impact on the output of the model. Pruning can simplify the model and improve its performance by reducing noise and focusing on the most relevant input features. Our results provide compelling evidence that our DL architecture is benefit from pruning technology. After pruning, the SHAP analysis and external test results indicate that pruned sparse model maintains the key input features and improves its generalization.

Our sparse model is capable of accurately identifying patients with different mutational profiles, which have significant implications for clinical management. Prior analysis methods can only associate mutational signatures with crude etiology. Via incorporating SHAP and gene accumulation analyses, our model not only can accurately predict metastasis risk in cancer patients, but for those who have metastasized, our approach can further stratify patients based on detailed molecular characterization with precision, portending favorable implications to inform personalized clinical management regimens. The capacity to segregate patients based on subtle genotypic distinctions enables tailored therapeutic interventions and prognostic predictions correlating with specific mutational profiles.

Further investigation will be justified to evaluate the clinical utility of this model, such as identifying actionable information from defined patient risk groups, identifying the significant relationship among mutations in coding and non-coding regions for metastasis. Due to the data confidentiality limitations, we did not conduct further research in this direction. It is important to investigate the interpretability of the features used by the model to predict prognosis for clinical guidance, which will be a topic of future work. In addition, having paired genomic data of the primary tumor and metastatic lesions from the same patients provides a powerful resource to study the genomic evolution of metastasis. Tracking the genomic changes from primary to metastatic sites in the same individual captures the trajectory of tumor evolution and progression in an authentic biological context, which will reveal core genomic features that distinguish metastatic clones from primary ones while eliminating the individual difference effects. Moreover, Martínez-Jiménez et. al. found that the metastatic tumors have a higher frequency of structural variants[8]. Incorporating additional structural mutation signature data and copy number variations (CNVs) has the potential to further improve AI algorithms from diverse multi-omics datasets. Leveraging diverse genomic data types, including structural variants, expression data across diverse populations worldwide will promote advances in AI for genomic medicines and enable more personalized therapies.

Methods

Genomic Data acquisition

The mutational signature contribution matrices pertaining to primary breast cancers within the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort, and metastatic breast cancers within the Hartwig Medical Foundation (HMF) cohort, were extracted from the supplementary tables of the reference[8]. The dataset encompasses two types of contribution matrices. The first one is denoted as ″signature contributions", encompassing the contributions for the mutational signatures detected in the two cohorts individually. The second type, termed ″etiology contributions", amalgamates the contributions of identical etiologies. For instance, contributions from mutational signatures SBS2 and SBS13 were conjoined to signify the collective contribution of the APOBEC etiology. In this study, we applied the ″signature contributions" matrices for the training and validation of our models.

To constitute an external test dataset, we retrieved somatic mutation data from the BRCA-EU cohort, encompassing primary breast cancers, via the International Cancer Genome Consortium (ICGC) data portal (https://dcc.icgc.org/). Furthermore, we sourced somatic mutation data from the POG570 cohort[49], containing metastatic breast cancers, from https://www.bcgsc.ca/downloads/POG570. In a bid to uphold data quality, samples characterized by low mutation burdens (mutation count < 50) were systematically excluded from the analysis. This curation process led to a refined dataset consist of 496 primary breast cancer samples and 127 metastatic breast cancer samples.

Mutational signatures extraction

As shown in Figure 1d, to ascertain the mutational signature contributions from an external dataset, we implemented a comparable mutational signatures analysis pipeline, as delineated in reference[8], albeit with certain modifications. In essence, we categorized somatic mutations within the external dataset into 96

SBS, 78 DBS, and 83 ID classes, as expounded upon in the earlier study[43]. The relative frequencies of each category within these channels were computed through utilization of the R package mutSigExtractor (https://github.com/UMCUGenetics/mutSigExtractor, v1.23). In order to ensure agreement with the mutational signatures employed in our model training and to mitigate potential bleeding effects stemming from the mutational signature extraction process, we refrained from conducting de novo mutational signatures extraction on the external dataset. Instead, we opted to leverage the 14 SBS, 5 DBS, and 9 ID mutational signatures previously identified within breast cancer instances within the PCAWG and HMF cohorts, employing these as reference signatures. Then, the fitToSignatures() function of mutSigExtractor, which employing a least square fitting algorithm, was applied to ascertain the individualized contributions of the aforementioned reference signatures within each sample from the external dataset. The matrices denoting these contribution values were subsequently employed as input features for evaluating the performance of our model.

Implementation of MetaWise 2.0

The updated MetaWise framework consists of two modules: the classification module and the pathogenic process identification module. The workflow is shown is Figure 1c. The classification module aims to predict the metastasis possibility of a patient based on their genomic data. The pathogenic process identification module aims to identify the key biological processes that are associated with metastasis process from the genomic profile of patients.

Model design and evaluation

We implemented our model using the Keras framework with Tensorflow as the backend. Our model consists of fully-connected layers, each followed by a batch normalization layer and a ReLU activation function, with a softmax output layer. To optimize the performance of our model, we fine-tuned various hyperparameters, such as the learning rate, the weight decay, the dropout rate, and the activation function. The pruning process was performed by Keras prune_low_magnitude function.

We evaluated the performance of the updated approach and compared with our previous model, MetaWise[17] by five-fold cross validation. We measured the accuracy, recall, precision, specificity, F1-score, the matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPR) on both internal and external test sets. We compared the results of weight pruning, input feature pruning and global pruning to identify the best pruning strategy.

Model interpretability analysis

We employed SHAP (SHapley Additive exPlanations) analysis to gain insights into the explanation of our model and to reveal the significance and impact of various mutation signatures on the prediction. SHAP analysis can offer both global and local explanations of the model, as well as feature interactions and dependencies. We utilized the Python library, shap[50], to conduct SHAP analysis for our model. After we identified the important mutation signatures for each sub-set data, we associated genes with mutational signatures using gene cumulative contribution abundance analysis[25] to understand the pathogenesis of patients.

Cumulative contribution abundance analysis

In order to in-depth mine the relationship between genes and mutational signatures, the mutational signature analysis was performed to associate mutational signatures with genes. A simple and practical R package, RNMF[25], was applied by analyzing cumulative contribution abundance of genes.

GO enrichment analysis

Gene Ontology (GO) term enrichment analysis was performed to identify over-represented GO terms in our gene set of interest. The GO system of classification assigns genes to a set of predefined bins based on their functional characteristics. Enrichment analysis identifies which GO terms are over-represented (or

under-represented) in the gene set using annotations for that gene set. Enrichment analysis was performed using the clusterProfiler[51], which is an R package that provides functions for statistical analysis and visualization of functional profiles for genes and gene clusters. It can be used to identify enriched gene sets in a cluster of genes, or to compare the functional profiles of two or more clusters. The GO aspect (molecular function, biological process, cellular component) for the analysis was selected, as well as the species from which the genes come. The results page displays a table that lists significant shared GO terms used to describe the set of genes entered.

Statistical analysis and results visualization

Ablation experiment

To perform the ablation experiments, we first ranked the input features according to their Shapley values. Then, we removed the top-ranked input features one at a time, retrained the model on the remaining features, and evaluated its performance using a variety of metrics. The results of these ablation experiments were analyzed to determine the impact of each removed input feature on model performance.

meta-analysis

Additional pathway enrichment analyses, gene list annotations and protein-protein interaction network analysis were performed using the free online meta-analysis tool Metascape[52] (https://metascape.org/).

Code availability

Code will be uploaded to the github repository (https://github.com/promethiume/MetaWise) once the paper has been conditionally accepted, and are available from the corresponding author on reasonable request during the manuscript review process.

Acknowledgements

We thank G. Peng, Y. Xin and L. Wei from the Innovation Center of StoneWise, AI. Ltd. for their helpful discussions and support; X.Xiang, S.Guo, Y.Wang and our colleagues at StoneWise, AI. Ltd. for their support and encouragement.

Author contributions statement

Y.Z. and M.P. initiated the project. M.P., X.L. and K.F. conducted the research. W.Z. and K.F. curated and pre-processed the data. M.P. designed the method. M.P., X.L, K.F. and K.T. performed the evaluation and analyzed the results. L.Z. contributed on data visualization. M.P., and K.T. wrote the manuscript. All authors reviewed manuscript.

Additional information

The authors declare no competing interests.

References

1. Lambert, A.W., Pattabiraman, D.R. & Weinberg, R.A. Emerging Biological Principles of Metastasis. *Cell* **168**, 670-691 (2017).

2. Massagué, J. & Obenauf, A.C. Metastatic colonization by circulating tumour cells. *Nature* **529**, 298-306 (2016).
3. Paul, M.R. *et al.* Genomic landscape of metastatic breast cancer identifies preferentially dysregulated pathways and targets. *The Journal of Clinical Investigation* **130**, 4252-4265 (2020).
4. Harbeck, N. *et al.* Breast cancer. *Nat Rev Dis Primers* **5**, 66 (2019).
5. Lefebvre, C. *et al.* Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. *PLOS Medicine* **13**, e1002201 (2016).
6. Nik-Zainal, S. & Morganella, S. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. *Clinical Cancer Research* **23**, 2617-2629 (2017).
7. Stephens, P.J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404 (2012).
8. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333-341 (2023).
9. IEEE Transactions on Biomedical Engineering (T-BME). *IEEE Transactions on Biomedical Engineering* **68**, C3-C3 (2021).
10. Windsor, G.O., Bai, H., Lourenco, A.P. & Jiao, Z. Application of artificial intelligence in predicting lymph node metastasis in breast cancer. *Frontiers in Radiology* **3**(2023).
11. Sella, N. *et al.* Interactive exploration of a global clinical network from a large breast cancer cohort. *npj Digital Medicine* **5**, 113 (2022).
12. Albaradei, S. *et al.* Machine learning and deep learning methods that use omics data for metastasis prediction. *Computational and Structural Biotechnology Journal* **19**, 5008-5018 (2021).
13. Cosgrove, N. *et al.* Mapping molecular subtype specific alterations in breast cancer brain metastases identifies clinically relevant vulnerabilities. *Nature Communications* **13**, 514 (2022).
14. Jiang, B. *et al.* Machine learning of genomic features in organotropic metastases stratifies progression risk of primary tumors. *Nature Communications* **12**, 6692 (2021).
15. Albaradei, S. *et al.* MetastaSite: Predicting metastasis to different sites using deep learning with gene expression data. *Frontiers in Molecular Biosciences* **9**(2022).
16. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications* **11**, 728 (2020).
17. Zheng, W. *et al.* Deep learning model accurately classifies metastatic tumors from primary tumors based on mutational signatures. *Scientific Reports* **13**, 8752 (2023).
18. Xu, Y., Cui, X. & Wang, Y. Pan-Cancer Metastasis Prediction Based on Graph Deep Learning Method. *Frontiers in Cell and Developmental Biology* **9**(2021).
19. Abdollahi, S., Lin, P.-C. & Chiang, J.-H. DiaDeL: An Accurate Deep Learning-Based Model With Mutational Signatures for Predicting Metastasis Stage and Cancer Types. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **19**, 1336–1343 (2021).
20. Snoek, J., Larochelle, H. & Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. in *Advances in Neural Information Processing Systems* Vol. 25 (eds Pereira, F., Burges, C.J., Bottou, L. & Weinberger, K.Q.) (Curran Associates, Inc., 2012).
21. Pfeifer, G.P. Mutagenesis at Methylated CpG Sequences. in *DNA Methylation: Basic Mechanisms* (eds. Doerfler, W. & Böhm, P.) 259-281 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).

22. Liu, Y., Gusev, A., Heng, Y.J., Alexandrov, L.B. & Kraft, P. Somatic mutational profiles and germline polygenic risk scores in human cancer. *Genome Med* **14**, 14 (2022).
23. Singh, V.K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Communications Biology* **3**, 421 (2020).
24. Wong, J.K.L. *et al.* Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions. *Nature Communications* **13**, 178 (2022).
25. Li, Z., Liang, H., Zhang, S. & Luo, W. A practical framework RNMF for exploring the association between mutational signatures and genes using gene cumulative contribution abundance. *Cancer Medicine* **11**, 4053-4069 (2022).
26. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31-46 (2022).
27. Gui, P. & Bivona, T.G. Evolution of metastasis: new tools and insights. *Trends Cancer* **8**, 98-109 (2022).
28. Martínez-Jiménez, F. *et al.* Genetic immune escape landscape in primary and metastatic cancer. *Nat Genet* **55**, 820-831 (2023).
29. Gote, V., Nookala, A.R., Bolla, P.K. & Pal, D. Drug Resistance in Metastatic Breast Cancer: Tumor Targeted Nanomedicine to the Rescue. *Int J Mol Sci* **22**(2021).
30. Fares, J., Fares, M.Y., Khachfe, H.H., Salhab, H.A. & Fares, Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct Target Ther* **5**, 28 (2020).
31. Santos, C.R. & Schulze, A. Lipid metabolism in cancer. *Febs j* **279**, 2610-23 (2012).
32. Chaudhary, P.K. & Kim, S. An Insight into GPCR and G-Proteins as Cancer Drivers. *Cells* **10**(2021).
33. SenGupta, S., Parent, C.A. & Bear, J.E. The principles of directed cell migration. *Nat Rev Mol Cell Biol* **22**, 529-547 (2021).
34. Papanicolaou, M. *et al.* Temporal profiling of the breast tumour microenvironment reveals collagen XII as a driver of metastasis. *Nat Commun* **13**, 4587 (2022).
35. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517-524 (2022).
36. Benzina, S. *et al.* Pax-5 is a potent regulator of E-cadherin and breast cancer malignant processes. *Oncotarget* **8**, 12052-12066 (2017).
37. Leblanc, N., Harquail, J., Crapoulet, N., Ouellette, R.J. & Robichaud, G.A. Pax-5 Inhibits Breast Cancer Proliferation Through MiR-215 Up-regulation. *Anticancer Res* **38**, 5013-5026 (2018).
38. Benzina, S. *et al.* Breast Cancer Malignant Processes are Regulated by Pax-5 Through the Disruption of FAK Signaling Pathways. *J Cancer* **7**, 2035-2044 (2016).
39. Dong, T. *et al.* FOXO1 inhibits the invasion and metastasis of hepatocellular carcinoma by reversing ZEB2-induced epithelial-mesenchymal transition. *Oncotarget* **8**, 1703-1713 (2017).
40. Storz, P., Döppler, H., Copland, J.A., Simpson, K.J. & Toker, A. FOXO3a promotes tumor cell invasion through the induction of matrix metalloproteinases. *Mol Cell Biol* **29**, 4906-17 (2009).
41. Feng, X. *et al.* Cdc25A regulates matrix metalloprotease 1 through Foxo1 and mediates metastasis of breast cancer cells. *Mol Cell Biol* **31**, 3457-71 (2011).

42. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* **21**, 619-637 (2021).
43. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
44. Thatikonda, V. *et al.* Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers. *Nat Cancer* **4**, 276-289 (2023).
45. Zagouri, F., Bournakis, E., Koutsoukos, K. & Papadimitriou, C.A. Heat shock protein 90 (hsp90) expression and breast cancer. *Pharmaceuticals (Basel)* **5**, 1008-20 (2012).
46. Tian, Y. *et al.* Extracellular Hsp90α and clusterin synergistically promote breast cancer epithelial-to-mesenchymal transition and metastasis via LRP1. *J Cell Sci* **132**(2019).
47. Stellas, D., El Hamidieh, A. & Patsavoudi, E. Monoclonal antibody 4C5 prevents activation of MMP2 and MMP9 by disrupting their interaction with extracellular HSP90 and inhibits formation of metastatic breast cancer cell deposits. *BMC Cell Biol* **11**, 51 (2010).
48. Jin, T., Carbin, M., Roy, D., Frankle, J. & Dziugaite, G.K. Pruning's effect on generalization through the lens of training and regularization. *Advances in Neural Information Processing Systems* **35**, 37947-37961 (2022).
49. Pleasance, E. *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat Cancer* **1**, 452-468 (2020).
50. Lundberg, S.M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. *et al.*) 4765-4774 (Curran Associates, Inc., 2017).
51. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
52. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications* **10**, 1523 (2019).