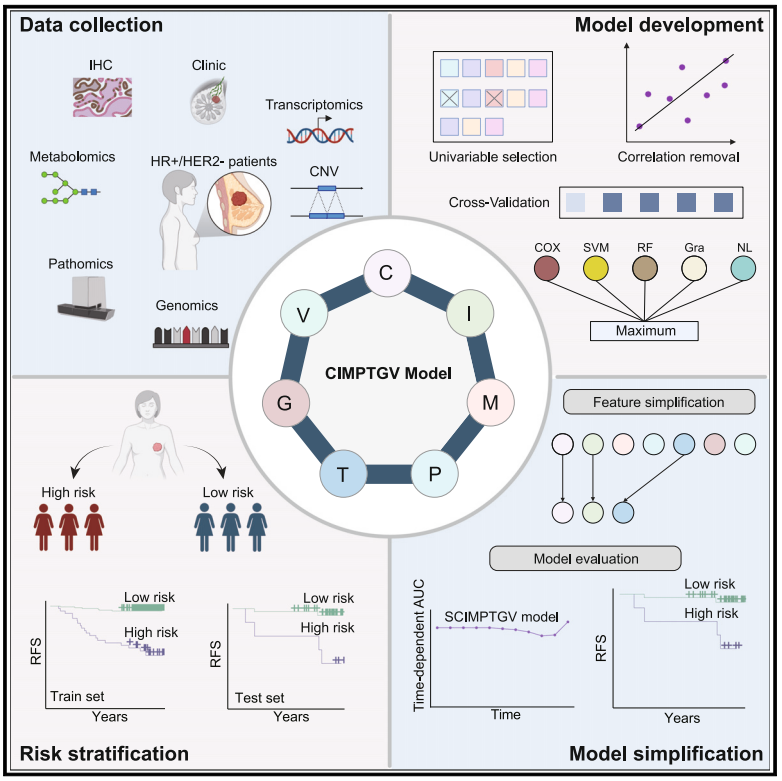
Cell Reports
Medicine

Article

Multimodal integration using a machine learning approach facilitates risk stratification in HR+/HER2− breast cancer

Graphical abstract



Highlights

- The CIMPTGV model is a multi-modality model generated by machine learning approach
- The CIMPTGV model predicts relapse in HR+/HER2− breast cancer patients
- Synergistic and complementary effects enhance efficacy of the CIMPTGV model
- A simplified CIMPTGV model balances efficacy with reduced data collection costs

Authors

Hang Zhang, Fan Yang, Ying Xu, Shen Zhao, Yi-Zhou Jiang, Zhi-Ming Shao, Yi Xiao

Correspondence

yizhoujiang@fudan.edu.cn (Y.-Z.J.), zhimingshao@fudan.edu.cn (Z.-M.S.), yixiao11@fudan.edu.cn (Y.X.)

In brief

Zhang et al. develop a machine-learning-based model, which integrates seven modalities to predict relapse in HR+/HER2− breast cancer patients. The model shows great efficacy and reveals biological insights. The simplified version of the model balances efficacy with reduced cost, offering a useful approach for clinical applications.

Zhang et al., 2025, Cell Reports Medicine 6, 101924
February 18, 2025
https://doi.org/10.1016/j.xcrm.2024.101924

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Multimodal integration using a machine learning approach facilitates risk stratification in HR+/HER2− breast cancer

Hang Zhang,[1,2] Fan Yang,[1,2] Ying Xu,[1,2] Shen Zhao,[1] Yi-Zhou Jiang,[1,*] Zhi-Ming Shao,[1,*] and Yi Xiao[1,3,*]
[1]Key Laboratory of Breast Cancer in Shanghai, Department of Breast Surgery, Fudan University Shanghai Cancer Center, Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, P.R.China
[2]These authors contributed equally
[3]Lead contact
*Correspondence: yizhoujiang@fudan.edu.cn (Y.-Z.J.), zhimingshao@fudan.edu.cn (Z.-M.S.), yixiao11@fudan.edu.cn (Y.X.)
https://doi.org/10.1016/j.xcrm.2024.101924

SUMMARY

Hormone receptor-positive (HR+)/human epidermal growth factor receptor 2-negative (HER2−) breast cancer is the most common type of breast cancer, with continuous recurrence remaining an important clinical issue. Current relapse predictive models in HR+/HER2− breast cancer patients still have limitations. The integration of multidimensional data represents a promising alternative for predicting relapse. In this study, we leverage our multi-omics cohort comprising 579 HR+/HER2− breast cancer patients (200 patients with complete data across 7 modalities) and develop a machine-learning-based model, namely CIMPTGV, which integrates clinical information, immunohistochemistry, metabolomics, pathomics, transcriptomics, genomics, and copy number variations to predict recurrence risk of HR+/HER2− breast cancer. This model achieves concordance indices (C-indices) of 0.871 and 0.869 in the train and test sets, respectively. The risk population predicted by the CIMPTGV model encompasses those identified by single-modality models. Feature analysis reveals that synergistic and complementary effects exist in different modalities. Simultaneously, we develop a simplified model with a mean area under the curve (AUC) of 0.840, presenting a useful approach for clinical applications.

INTRODUCTION

Breast cancer has now emerged as one of the most prevalent cancer among women.[1] Hormone receptor-positive (HR+)/human epidermal growth factor receptor 2-negative (HER2−) breast cancer constitutes the predominant subtype, encompassing 65%–70% of breast cancers.[2] As a pivotal clinical concern, the recurrence of HR+/HER2− breast cancer is a continuous process, characterized by an annual distant recurrence rate held at 1%. Up to 20% of HR+/HER2− breast cancer patients experienced recurrent metastases[3] due to endocrine-resistant mechanisms.[4–6] Even after a period of 5–10 years of standard endocrine therapy, recurrence persists steadily.[7,8] Consequently, there arises an imperative need to predict the risk of recurrence in HR+/HER2− breast cancer patients.

Advancements in biomarker discovery, multi-omics technology, and optimization of prognostic prediction models enable the prediction of treatment resistance and recurrence risk.[9–12] Gene expression profiling, such as the Oncotype DX assay and MammaPrint, provides strategies to stratify patients into low-risk and high-risk groups and helps identify those who might benefit from chemotherapy in early-stage ER+ breast cancer patients.[13,14] However, these predictive signatures exhibit limitations in certain scenarios, evidenced by their modest performance in predicting recurrence in node-positive breast cancer, with concordance indices (C-indices) ranging from 0.56 to 0.63.[15] This limitation is attributed, in part, to the predominant reliance on clinicopathological or transcriptomic data, which fail to comprehensively capture the intricacies of cancer.

With advancements in sequencing technology and decreasing costs, the multi-omics data, encompassing genome, transcriptome, proteome, metabolome, and epigenome, have become increasingly available. Moreover, the emergence of artificial intelligence (AI) technology has facilitated the integration of data from various modalities. In recent years, multimodal machine learning has been employed for treatment efficacy prediction, postoperative stratification, and prognostic assessment,[16–19] demonstrating its considerable practical utility. The integration of multiple modalities provides a comprehensive understanding of tumor characteristics in all directions, and machine learning approaches can transform these data into advanced prognostic models.[20] However, the integration of diverse omics types remains challenging due to issues related to data collection, cohort establishment, and integration methodologies.[21] Its application in breast cancer remains underexplored.

In this study, leveraging our multi-omics HR+/HER2− breast cancer cohort, we constructed a machine learning framework and generated a multimodal model named CIMPTGV. This model integrates clinical information, immunohistochemistry (IHC) data, transcriptomics, metabolomics, genomics, copy-number

Table 1. Clinical characteristic of patients in the cohort

	Overall (*N* = 547)
Age (year)	
Mean (SD)	52.7 (11.0)
Median [Min, Max]	52.0 [24.0, 90.0]
BMI (kg/m^2)	
Mean (SD)	23.6 (3.23)
Median [Min, Max]	23.4 [14.3, 47.7]
Menopause	
No	243 (44.4%)
Yes	304 (55.6%)
pT	
pT1	233 (42.6%)
pT2	305 (55.8%)
pT3	9 (1.6%)
pN	
pN0	226 (41.3%)
pN1	173 (31.6%)
pN2	86 (15.7%)
pN3	62 (11.3%)
AJCC stage	
I	139 (25.4%)
II	258 (47.2%)
III	150 (27.4%)
Vessel invasion	
No	247 (45.2%)
Yes	300 (54.8%)
RFS	
Not occurred	397 (72.6%)
Occurred	150 (27.4%)
OS	
Not occurred	472 (86.3%)
Occurred	75 (13.7%)
DMFS	
Not occurred	412 (74.9%)
Occurred	135 (25.1%)
Follow-up duration (month)	
Median [Q1, Q3]	79.1 [72.1, 93.1]

Abbreviations: BMI, body mass index; pT, pathologic tumor stage; pN, pathologic nodal stage; AJCC, American Joint Committee on Cancer; RFS, relapse-free survival; OS, overall survival; DMFS, distant metastasis-free survival.

variations (CNVs), and pathomics. The CIMPTGV model demonstrated great efficacy in predicting recurrence risk, with C-indices of 0.871 in the train set and 0.869 in the test set. Notably, the homologous recombination deficiency (HRD) score exhibited a significant positive correlation with the CIMPTGV risk score. Subsequent analysis revealed synergistic and complementary effects between different modalities, contributing to the high C-index of the CIMPTGV model. To promote clinical application, we also developed a simplified version of the CIMPTGV model. This simplified model outperforms common modality combinations in the clinical settings and holds promise for improving the prediction of HR+/HER2− breast cancer recurrence, facilitating the identification of high-risk patients and enhancing their prognosis through personalized treatment strategies.

RESULTS

Cohort development and machine learning framework construction

Our study included 579 patients diagnosed with HR+ (estrogen receptor-positive [ER+] or progesterone receptor-positive [PR+]) and HER2− unilateral invasive breast carcinoma between 2009 and 2016. Patients with distant metastasis or those who had received pre-treatment at the time of diagnosis were excluded (Figure S1). Among the participants, 547 patients had complete clinical and follow-up information, with a median follow-up duration of 79.1 months (interquartile range [IQR]: 72.1–93.1 months). During the follow-up period, 75 patients died of the disease, 150 experienced recurrence, and 135 experienced distant metastasis (Table 1).

The cohort consists of multimodal data including clinical information (*N* = 547), IHC data (*N* = 510), transcriptomics data (*N* = 565), metabolomic data (*N* = 380), genomic data (*N* = 467), CNV data (*N* = 429), and pathomic data (*N* = 418) (Figure 1A).

To fully utilize the multimodal data, we implemented a machine learning framework to integrate features into a predictive model. The framework consisted of four parts (Figure 1B). First, we established the train set and test set. To maximize the utility of our multi-omics data and improve predictive efficacy, we performed stratified sampling in each model cohort, using recurrence status as the stratification factor to divide the train set and test set in a 4:1 ratio. Subsequently, feature extraction was conducted in each modality within the model. The third part is removing features with correlations higher than 0.9 and selecting the top 15 features with the highest univariate Cox hazard analysis score to obtain the filtered feature set. The final step involved model training and testing. The filtered feature set was input into five models, including Cox's proportional hazard's model, survival support vector machine, random survival forest model, DeepSurv non-linear model, and gradient boosting survival model. Hyperparameters were optimized using a randomized 1,000-step 5-fold cross-validation search to maximize the average cross-validation C-index, and the model with the highest average cross-validation C-index was selected as the final model. The final model was tested on the aforementioned hold-out test set to evaluate its performance.

Multimodal integration improves the prediction efficacy and benefits risk stratification

Through the machine learning framework, we developed models with varying numbers of modality dimensions, ranging from a single modality to complete 7 modalities. In both the train and test sets, the models' C-index progressively increased with the number of modality dimensions increased, demonstrating statistically significant improvements (Figures 2A and 2B). Notably, in the train set, the CIMPTGV model, which integrated all seven modalities, achieved the highest C-index of 0.871. Similarly, in the test set, the CIMPTGV model also achieved a

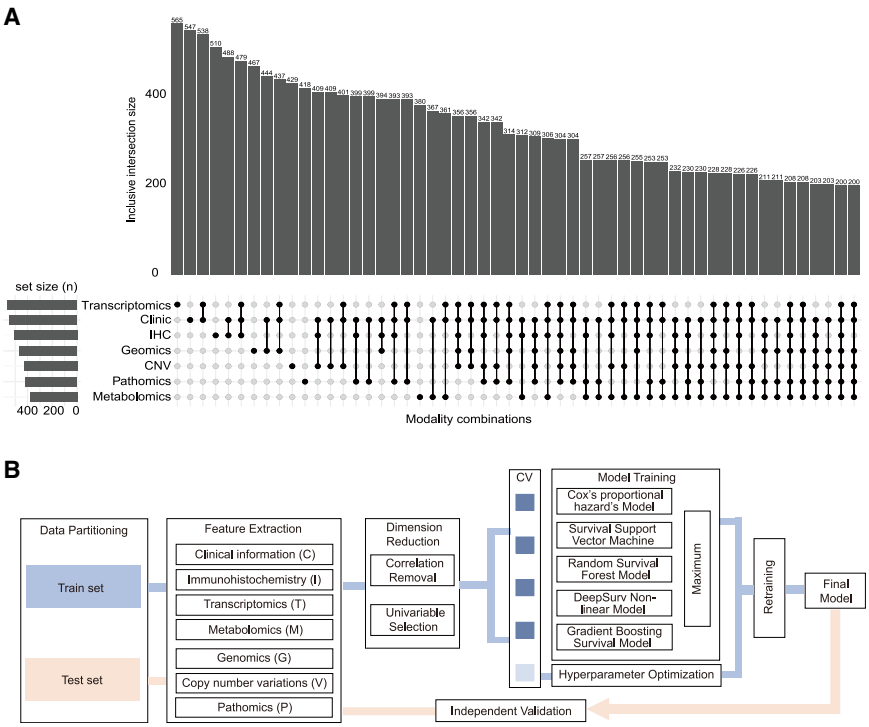


Figure 1. Cohort development and machine learning framework construction

(A) Upset plot illustrated the inclusive intersection sizes for different combinations of modalities. The vertical bars in the upper plot represented the number of patients in each modality combination, indicated by the black circles in the plot located below. The modalities included clinical information (N = 547), immunohistochemistry data (N = 510), transcriptomics data (N = 565), metabolomics data (N = 380), genomics data (N = 467), copy-number variation data (N = 429), and pathomics data (N = 418). The set sizes vary from 200 to 565.

(B) The machine learning flowchart consists of steps including data partitioning, feature extraction, dimension reduction, model training, and independent validation. The train and test sets varied according to the specific combinations of modalities used.

Abbreviations: IHC, immunohistochemistry; CNV, copy-number variation; CV, cross-validation.

high C-index of 0.869, outperforming all other models assessed. Furthermore, the CIMPTGV model exhibited superior predictive efficacy compared to traditional modality combinations, such as CIT (clinical information, immunohistochemistry information, and transcriptomics), CT (clinical information and transcriptomics), and CI (clinical information and immunohistochemistry information) models, which have been applied in HR+/HER2− breast cancer.[22–24]

By applying the CIMPTGV model, we generated recurrence risk prediction scores for both the train and test sets. By setting the upper quartile of the predicted risk score of the train set as the cutoff value, we stratified patients into high-risk and low-risk groups. It was noticed that the CIMPTGV model effectively distinguished relapse-free survival (RFS), overall survival (OS), and distant metastasis-free survival (DMFS), in both the train and test sets (Figure 2C).

To further validate the stability of the CIMPTGV model, we randomly divided the population with the full modality into train and test sets 100 times. We then retrained and retested the CIMPTGV model, recording the average cross-validation C-indices of the train set and the C-indices of the test set. The results showed that the C-indices for both the train and test sets mostly ranged from 0.8 to 0.9, with differences less than 0.05 (Figure S2A). This indicates that the current training method is stable and that the CIMPTGV model is not a result of randomness. Additionally, by iterating through the model's hyperparameters, we observed that increasing the number of estimators initially improved both training and testing performance, reaching an optimal point. However, further increases in the number of estimators resulted in overfitting. The number of estimators we selected through cross-validation hyperparameter tuning was 10, which is actually the optimal point, indicating that there was no overfitting in the current hyperparameter setting (Figure S2B).

Furthermore, we compared the CIMPTGV model with commercial assays such as Oncotype DX and MammaPrint. We selected an overlapping patient cohort from the CIMPTGV cohort and the patients applicable to Oncotype DX (or MammaPrint). For the Oncotype DX, we used RNA expression of genes as a substitute for node-negative patients or postmenopausal patients with lymph node metastasis (n = 133). For the MammaPrint score, we included clinically high-risk elderly patients (n = 94) and input the 70 genes into Cox's proportional hazards model, using its predicted risk score as the MammaPrint substitute score. The results revealed that the CIMPTGV model outperformed both the MammaPrint Cox score (C-index: CIMPTGV 0.812, 95% confidence interval: 0.758–0.866 vs. MammaPrint Cox score 0.688, 95% confidence intervals: 0.659–0.716) and the Oncotype RNA score (C-index: CIMPTGV 0.866, 95% confidence intervals: 0.830–0.900 vs. Oncotype RNA score 0.568, 95% confidence intervals: 0.477–0.627) (Figure S2C). Additionally, the CIMPTGV model significantly distinguished patients in the high-risk group for recurrence or death. This model adeptly identified 74.2% (23/31) of the patients who experienced recurrence and predicted an 85.8% (145/169) probability of the low-risk group remaining relapse-free during the follow-up period, emphasizing its robust predictive capability (Figure S2D and Table S1).

To further compare the predictive efficacy of the CIMPTGV model with other single-modality models, we evaluated the performance of each model using univariate Cox regression (Figure 2D). We found that the CIMPTGV model exhibited the highest and most statistically significant hazard ratio score. Multivariate Cox regression analysis further revealed the independent predictive efficacy of the CIMPTGV model (Figure S2E). Subgroup analysis indicated that the CIMPTGV model effectively distinguished high-risk recurrence populations across different stages

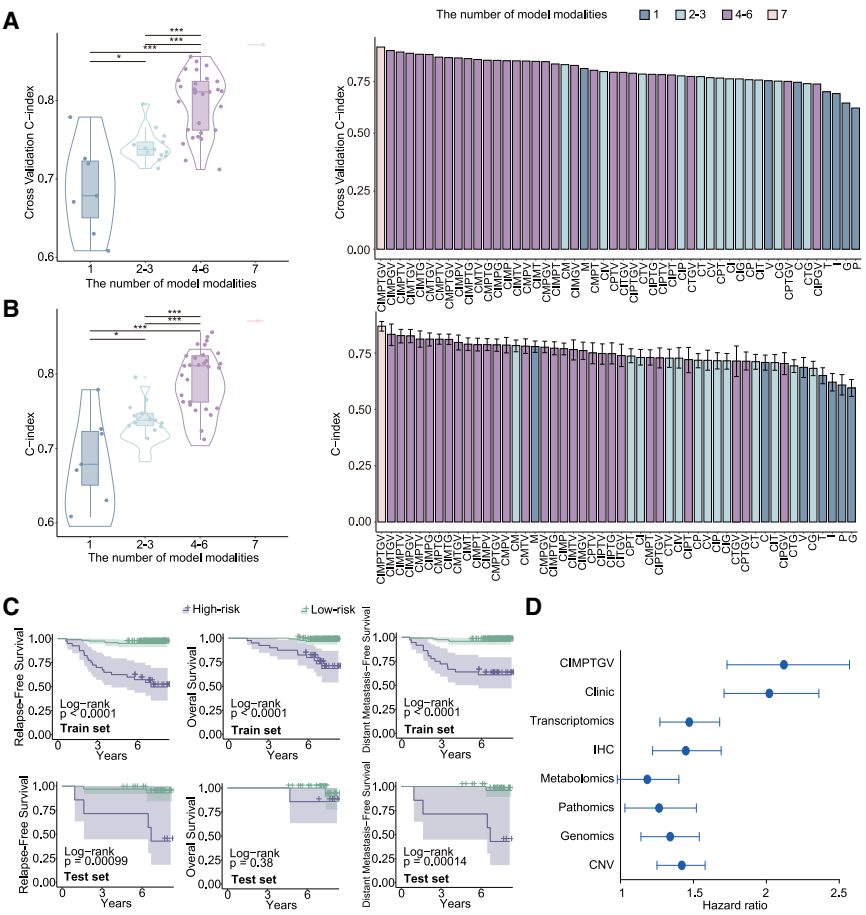


Figure 2. Multimodal integration improves the prediction efficacy and benefits risk stratification

(A and B) The C-indices of models combining different modalities were presented for both the train set (A, left) and the test set (B, left). Different model types with different numbers of modalities were distinguished by color. Comparisons of C-indices between models with different numbers of modalities (right) were shown. The interior boxes represented the 25th, 50th, and 75th percentiles. The bars indicated the C-indices of each model, and the lower and upper points of the error bars in the test set represented the 95% confidence intervals, derived from 1,000-fold bootstrapping. Dots indicated individual C-index, and the colors differentiated between model types (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, p values were obtained from a Mann-Whitney U-test).

(C) Kaplan-Meier survival analyses were conducted to compare RFS, OS, and DMFS between the high-risk and low-risk groups stratified by the CIMPTGV model in both the train and test sets. The p value was obtained from the log rank test.

(D) Hazard ratio with 95% confidence intervals for each model in the test set calculated by using a univariate Cox regression analysis.

Abbreviations: IHC, immunohistochemistry; CNV, copy-number variation; Patho, pathomics; C-index, concordance index; RFS, relapse-free survival; OS, overall survival; DMFS, distant metastasis-free survival.

and different clinical or IHC characteristics in breast cancer patients (Figure S2F). These findings highlight the robust capability of the CIMPTGV model to predict survival prognosis, demonstrating its capability to independently predict RFS beyond clinical variables.

Orthogonal data exist in multiple modalities, thus improving predictive efficacy

Modalities with orthogonal information enhance predictive efficacy, whereas modalities with mutual information do not achieve the same improvement.[20] To evaluate the predictive associations of different dimensions within the CIMPTGV model, we trained single-modality models on the train set of the CIMPTGV model using features from individual dimensions and evaluated these single-modality models using the test set of the CIMPTGV model. We analyzed the correlation of predictive scores from each single-modality model (Figure 3A). Our analysis revealed a low correlation in predictive scores between single-modality models, with most modalities showing the absolute values of Pearson correlation coefficients no more than 0.3. By setting the median predicted risk score of the train set as the cutoff value for both the train and test sets, significant disparities were observed in the predicted risk populations among different single-modality models (Figure 3B), highlighting the existence of orthogonal information between different modalities. Besides, correlation analysis between features from different modalities (n = 7) within the CIMPTGV model revealed weak associations. The average feature expression analysis also confirmed the low inter-modality correlation, indicating the orthogonality between modalities (Figure S3). This finding emphasizes that a singular dimension is insufficient to comprehensively illustrate the biological characteristics of patients, contributing to the suboptimal predictive efficacy of single-modality models. In contrast, multi-modality models effectively predict patient prognosis by integrating data from different modalities. The integrated orthogonal data augment information beyond any other single-modality data.

A heatmap of predictive scores further validated our findings, demonstrating that the CIMPTGV risk population encompassed the high-risk population identified by other single-modality models, with the CIMPTGV risk population exhibiting a higher recurrence ratio in the test set (Figure 3C). For the extremely low-risk population identified by the CIMPTGV model, almost all other single-modality dimensions also indicated low risk. Additionally, as the number of single-modality dimensions with higher predicted risk scores increases, the predicted risk score in the CIMPTGV model also increases, indicating that the CIMPTGV model effectively leverages the orthogonal information from other single-modality dimensions to enhance its predictive performance.

CellPress
OPEN ACCESS

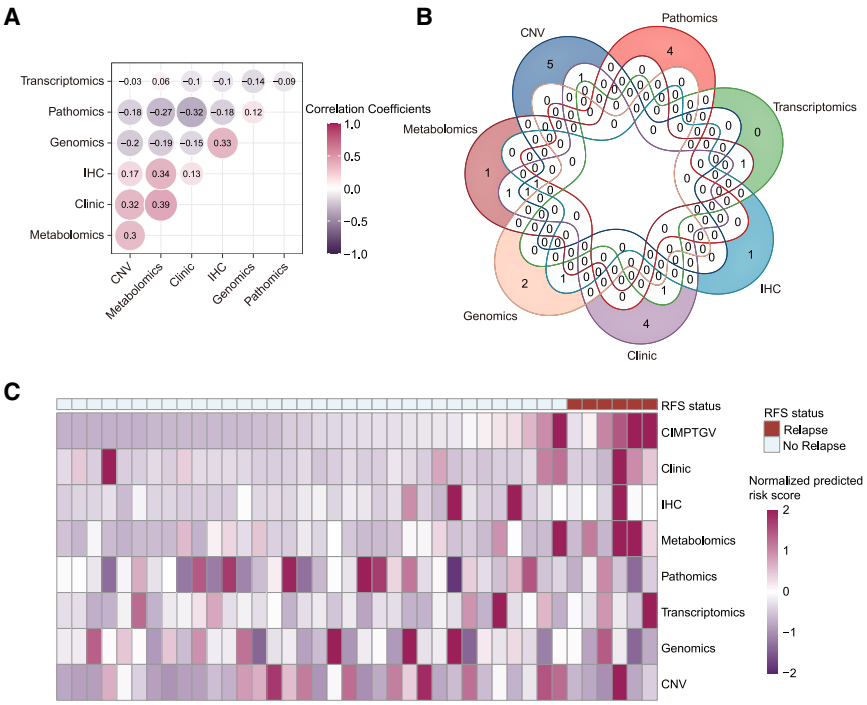


Figure 3. Orthogonal data exist in multiple modalities, thus improving predictive efficacy

(A) Pearson correlation coefficients of the prediction scores of each single-modality model in the test set.

(B) A Venn diagram comparing the high-risk patients identified by each single-modality model in the test set. For each single-modality model, individual predictive scores were generated, and patients with predictive scores above the cutoff score were selected as risk populations. The number of common risk populations between different models was indicated in the figure.

(C) Heatmap showing the scaled predicted risk score from the CIMPTGV model and other single-modality models in the test set, with RFS status annotated above.

Abbreviations: IHC, immunohistochemistry; CNV, copy-number variation; RFS, relapse-free survival.

Manifestation of features of each modality in the CIMPTGV model

To comprehensively illustrate the characteristics of the high-risk and low-risk groups, we delineated the features incorporated in the CIMPTGV model (Figure 4A). As annotated, the high-risk group predicted by the CIMPTGV model in the whole cohort, including both the train and test sets, encompassed nearly all patients with recurrence. Consistently, patients in the high-risk group exhibited higher tumor stages, lymph node stages, and Ki-67 proliferation indices (Figure 4B). Among other dimensions, metabolomics and transcriptomics data exhibited the most notable distinctions between the high-risk and low-risk groups. Several nucleic acid metabolites, including pseudouridine, N4-Acetylcytidine, and 5′-methylthioadenosine, were significantly enriched in the high-risk group (Figure S4A). Conversely, some lipids, such as eicosadienoic acid, resolvin D1, and linoleic acid, were downregulated in the high-risk group (Figure S4B). Similarly, in the transcriptomic dimension, multiple pathways were differentially expressed. The Myc proto-oncogene (*MYC*) target pathway was highly expressed in the high-risk group, while the fatty acid metabolism pathway exhibited reduced expression in the high-risk group (Figure 4B).

Pathologically, the high-risk group exhibited a higher proportion of tumor cells (Figure 4B). Pathological topological features showed elevated levels of tumor cell betweenness in the high-risk group, indicating shorter spatial distances between tumor cells (Figure S4C). This reflects a greater degree of tumor cell clustering in the high-risk group, which is associated with a higher recurrence rate. We also assessed the importance score of each modality by removing each modality data and using the reduced C-index as the importance score. The results indicated that metabolomics data and clinical information are relatively critical among all modalities, while digital pathology data and genomic features have a comparatively smaller impact (Figure S4D). Overall, these findings suggested a robust association between metabolic alterations and tumor recurrence in HR+/HER2− breast cancer patients.

Feature and modality correlation analysis further supports the existence of complementary information

To identify features significantly positively correlated with the CIMPTGV predicted risk score, we conducted a correlation analysis between feature values and CIMPTGV predicted risk scores (Figure S5A). We found a significant positive correlation between the HRD score and the CIMPTGV predicted risk score, with the high-risk population exhibiting an elevated HRD score (Figure 5A). HRD can lead to genomic instability, ultimately promoting tumor progression.[25] Notably, we observed a substantial correlation between HRD and features from different dimensions. The HRD score was significantly positively correlated with the amplification of chromosome 11q13.3 (Figure 5B). Fibroblast growth factor 3 (*FGF3*), fibroblast growth factor 4 (*FGF4*), and cortactin (*CTTN*), located in this focal amplification peak and known to be associated with malignant biological behaviors in cancer,[26–31] were markedly amplified in the high-HRD-score group (Figure 5C) and high-risk group (Figure S5B), showing cis-effects at the mRNA level (Figure 5D). At the pathological level, patients with a high HRD score exhibited a higher degree of tumor cell contact and a greater number of tumor cells (Figure 5E), as evidenced by pathological images (Figure S5C). In our previous research, we quantified the morphological intratumor heterogeneity (MITH) by computing the mean value of the standard deviation of morphological and texture features among tumor nuclei for each sample.[32] A comparison of MITH scores between the high-HRD-score group and the low-HRD-score group revealed increased pathological heterogeneity in the high-HRD-score group (Figure 5F), which is consistent

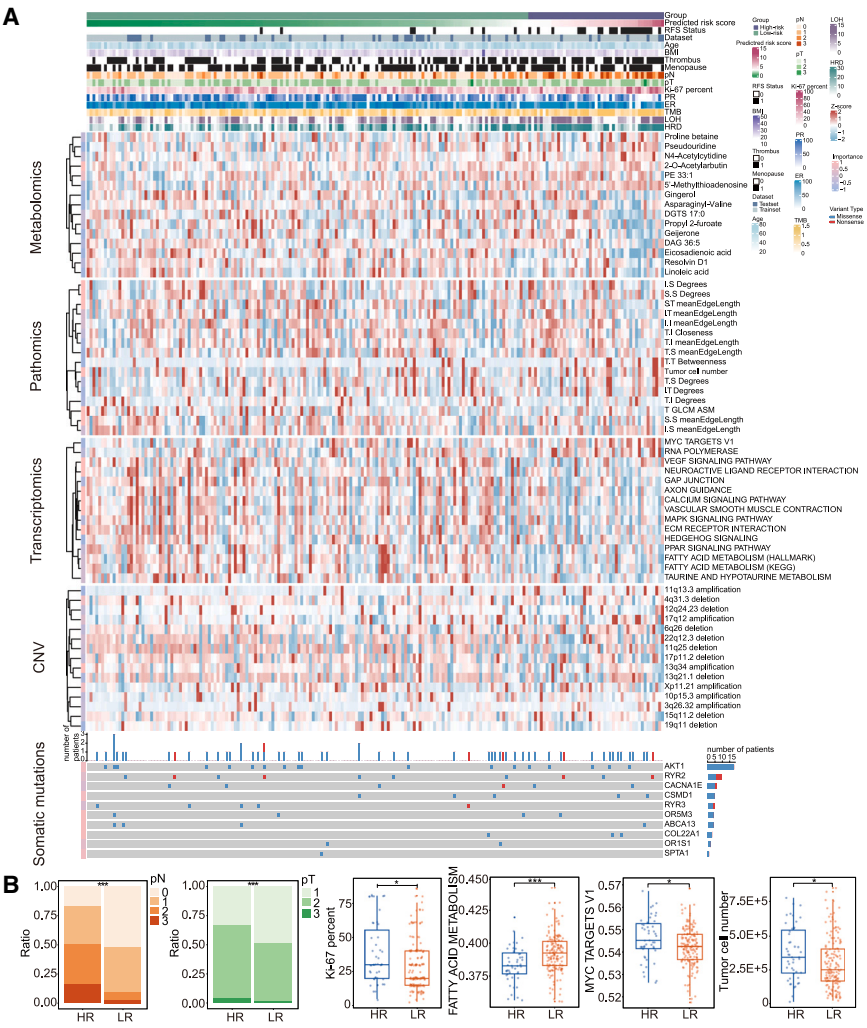


Figure 4. Manifestation of features of each modality in the CIMPTGV model

(A) Heatmap illustrating the *Z* scores of multimodal features used in the CIMPTGV model, with patients arranged according to the predicted risk scores generated by the model. Clinical, IHC, and selected genomics features were annotated above the heatmap, while metabolomics, pathomics, transcriptomics, and CNV were displayed within the heatmap. Somatic mutations of genes were displayed beneath the heatmap.

(B) Comparisons of pN, pT, Ki-67 proliferation index, fatty acid metabolism pathway, MYC target pathway, and tumor cell number between risk groups (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, p value was obtained from the chi-square test and the Mann-Whitney *U*-test).

Abbreviations: BMI, body mass index; TMB, tumor mutational burden; LOH, loss of heterozygosity; HRD, homologous recombination deficiency; GLCM, gray-level co-occurrence matrix; ASM, angular second moment of the co-occurrence matrix; I, immune cell; T, tumor cell; S, stroma cell; *MYC*, Myc proto-oncogene; HR, high risk; LR, low risk.

with the elevated recurrence risk in the high-HRD-score group. At the metabolomic level, patients with high HRD scores exhibited high enrichment of nucleic acid metabolites, indicating a close association between the genomic instability and tumor metabolic dysfunction (Figure 5G).

While patients with high HRD scores were at increased risk of recurrence, a subset of patients with low HRD scores also demonstrated high-risk recurrence (Figure S5D). Further analysis of other dimensions revealed complementary expression of features within this subgroup. These individuals exhibited higher clinical stages, certain enriched nucleic acid metabolites, and a greater average distance between tumor cells and immune cells (Figure S5E). Importantly, these features exhibited minimal correlation with the HRD score in the overall population (Figure S5F), indicating their independence from HRD-related influences. This observation underscores the presence of complementary information between different modalities.

In conclusion, the findings suggested that the enhanced predictive efficacy of CIMPTGV may be attributed to the synergistic and complementary effects arising from the integration of diverse modalities.

Simplified model construction enhances clinical applicability

To further enhance clinical applicability of the model, we developed a simplified multimodal model. We selected features from each dimension, incorporating all features from clinical, IHC, and pathomics dimensions due to their accessibility. For other dimensions, we included features with high feature importance scores in each modality (Figure 6A). The simplified model comprised basic patient information (age, body mass index[BMI], and menopausal status), clinical stage (pN stage and pT stage), IHC data (ER and PR percentage, Ki-67 proliferation index), digital pathological features, and features from other dimensions (transcriptomics, metabolomics, genomics, and CNVs) (Figure 6B). To verify the presence of orthogonal information within the simplified model, we constructed unimodal models using the corresponding features of each of the seven dimensions within the simplified model. For each unimodal model, we use the upper quartile of the predicted scores as the threshold to divide patients into high-risk and low-risk groups (Figure S6A). We found that, as the number of unimodal models predicting high-risk increased, the prognosis for these patients worsened. This finding suggested that information from different modalities is complementary, leading to a synergistic effect (Figure S6B). Furthermore, when the features of any single modality were removed from the simplified model, there was a significant decrease in its predictive performance (Figure S6C). This further underscores the importance of incorporating multimodal information to maintain the predictive efficacy of the simplified model. Although the predictive efficacy of the simplified model (S-CIMPTGV, mean area

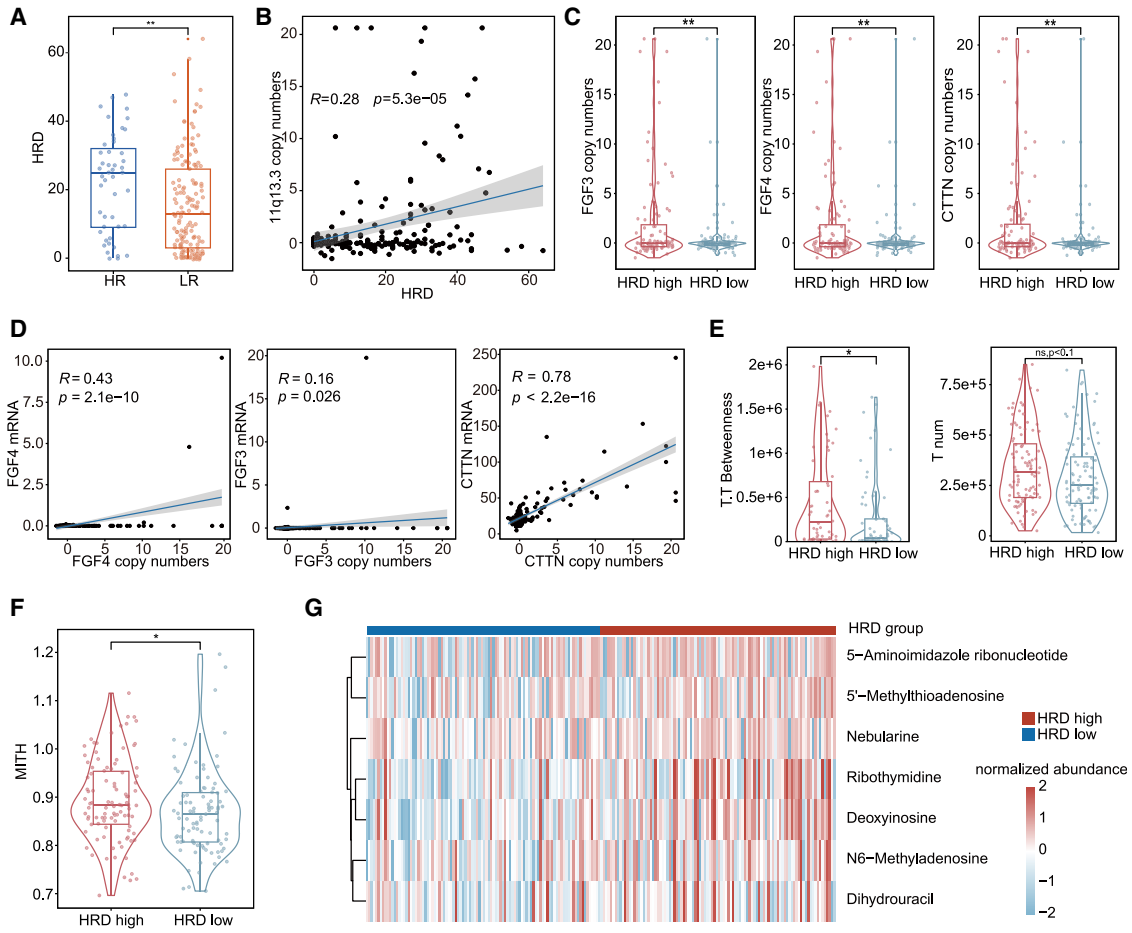


Figure 5. Modality correlation analysis further supports the existence of complementary information

(A) Comparison of HRD scores between the high-risk and low-risk groups stratified by the CIMPTGV model. Dots depicted individual HRD score, and the colors indicated different groups (**$p < 0.01$; the p value was obtained from a two-sided t test).

(B) Pearson correlation (two-sided) between the 11q 13.3 copy numbers and HRD score, with shaded regions representing 95% confidence intervals. The p value was calculated using a two-sided t test.

(C) Comparison of somatic copy numbers of three individual genes located at 11q13.3 between the high- and low-HRD-score groups, stratified by using median HRD score as cutoff (**$p < 0.01$; the p value was obtained from a two-sided t test).

(D) Pearson correlation (two-sided) between the somatic copy numbers and gene mRNA levels of three individual genes located at 11q13.3. Shaded regions represented 95% confidence intervals. The p value was calculated using a two-sided t test.

(E and F) Comparison of the spatial betweenness of tumor cells, the number of tumor cells, and MITH scores between the high- and low-HRD-score groups stratified by using median HRD score as cutoff (*$p < 0.05$; the p value was obtained from a two-sided t test).

(G) Comparison of the abundance of nucleic acid metabolites in the high- and low-HRD groups (*$p < 0.05$; the p value was obtained from a two-sided t test).

Abbreviations: HRD, homologous recombination repair; *FGF3*, fibroblast growth factor 3; *FGF4*, fibroblast growth factor 4; *CTTN*, cortactin; MITH, morphological intratumor heterogeneity; HR, high risk; LR, low risk.

under the curve [AUC] = 0.840) was slightly lower than that of the CIMPTGV model (mean AUC = 0.886), the S-CIMPTGV model outperformed other commonly used dimensions in clinical practice, as depicted by a relatively high AUC score over time (Figure 6C). Kaplan-Meier survival curves further illustrated significant stratification of predicted high- and low-risk groups, highlighting the utility of the simplified model in clinical decision-making (Figure 6D).

In conclusion, the simplified model streamlines the model-building process, contributing to its feasibility and effectiveness in clinical application.

DISCUSSION

The recurrence of HR+/HER2− breast cancer remains a significant clinical challenge. Even with limited tumor burden at diagnosis, the risk of disease recurrence and death persists after 5 years of standard endocrine therapy.[3,7,33] Although numerous prognostic molecules and signatures have been developed and tested and some of them already been applied in clinical settings,[13,14,34–37] the prognostic power is still limited in certain cases.[15] Utilizing our multi-omic data and machine learning approach, we developed a multi-modality model

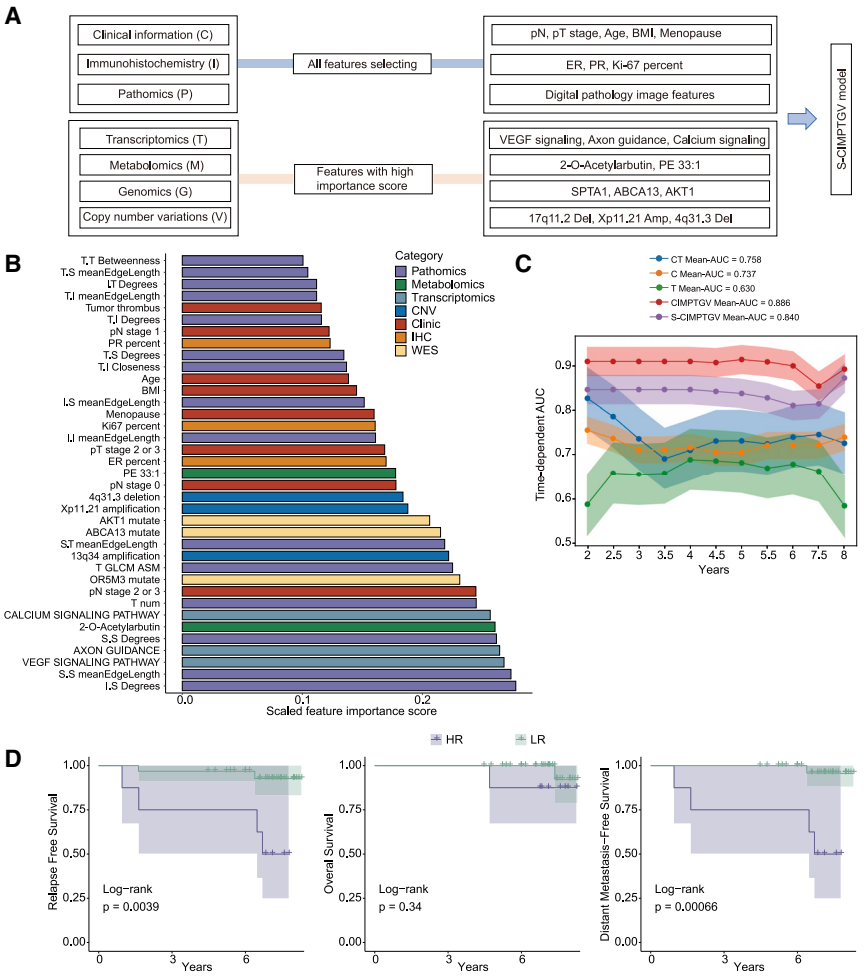


Figure 6. Simplified model construction promotes clinical application

(A) The workflow for constructing the simplified CIMPTGV model. Initially, all features from clinical information, immunohistochemistry and pathomics dimensions were incorporated. High-importance features from these dimensions were selected, and the filtered feature matrix was input into the machine learning framework to generate the predictive score.

(B) The scaled feature importance scores of features from different categories used in the simplified CIMPTGV model. Features were ranked by importance scores, with distinct feature categories represented by different colors.

(C) The area under the curve (AUC) of the time-dependent receiver operating characteristic (ROC) curve was computed for five models: CT (blue), C (orange), T (green), CIMPTGV (red), and the simplified CIMPTGV (purple) model. The mean AUC for each model was presented, with distinct colors used to differentiate the models.

(D) Kaplan-Meier survival analyses comparing RFS, OS, and DMFS between the high-risk and low-risk groups stratified by the S-CIMPTGV model in the test set. The *p* values were obtained from the log rank test.

Abbreviations: BMI, body mass index; ER, estrogen receptor; PR, progesterone receptor; *VEGF*, vascular endothelial growth factor; *SPTA1*, spectrin alpha, erythrocytic 1; *ABCA13*, ATP binding cassette subfamily A member 13; *AKT1*, AKT serine/threonine kinase 1; S-CIMPTGV, simplified CIMPTGV; AUC, area under the curve; CT, model integrated clinical information and transcriptomics; HR, high risk; LR, low risk.

named CIMPTGV. This model integrates clinical, IHC, metabolomics, pathomics, transcriptomics, genomics, and CNV data, achieving a C-index of 0.869 (Figures 1 and 2). The CIMPTGV model demonstrated excellent efficacy in discriminating risk groups for RFS, OS, and DMFS (Figure 2). We observed that CIMPTGV risk population encompassed the highest-risk population in other single-modality models (Figure 3). By delineating the characteristics of the CIMPTGV model, we observed enriched nucleic acid metabolites and higher levels of tumor cell aggregation in the high-risk group (Figure 4). Further analysis suggested that the robust predictive performance of the CIMPTGV model may be attributed to the existence of orthogonal and complementary information across different dimensions (Figure 5). Finally, the simplified model we established achieved a high C-index, effectively stratifying risk patients and simplifying the model construction process to promote clinical application (Figure 6).

It is known that AI-based methods digitize image data and utilize AI technologies to extract intricate information from images, holding significant potential for assisting radiologists and exploring image information. By utilizing routinely collected clinical data, such as pathological slices or radiographic images, AI not only facilitates early tumor screening[38–40] but also enhances diagnostic accuracy in conjunction with radiology.[41,42] It can predict biomarkers[43] and differentiate high-risk patients.[44] The relatively low cost of AI-based imaging studies and the capacity to utilize existing clinical data highlight its potential for widespread clinical application. However, traditional AI-based image-related research often focuses on clinical, pathological, or radiomic data, potentially neglecting other critical dimensions such as transcriptomics, metabolomics, and genomics.

Multimodal machine learning has emerged as a promising strategy for improving patient risk stratification. The integration of multiple modalities compensates for the limitations of single-modal understanding of tumor biology[20] and has been used in therapy response and survival prediction. For instance, a multimodal machine learning model for breast cancer undergoing neoadjuvant treatment predicted pathological complete response with an AUC of 0.87.[19] Similarly, a multimodal approach combining medical imaging, histopathological, and genomic features outperformed unimodal measures in predicting immunotherapy response in non-small cell lung cancer patients, achieving an AUC of 0.80.[45] However, limited by the scarcity of cohorts with multimodal information, the integration of multidimension data to guide clinical decision-making remains in its preliminary stage. These multi-omics data are rarely integrated even

when available, and few studies computationally exploit the prognostic potential of large-scale, multimodal integration.[20] Here, we established a large-scale multi-omics cohort of 579 Chinese HR+/HER2− breast cancer patients[46] and developed a model named CIMPTGV that combines all the modalities we used. The CIMPTGV model outperformed other models and showed great stratification efficacy. Although integrating multi-omics data presents challenges for clinical adoption due to its substantial cost, our simplified model effectively simplifies features. This simplification facilitates the design of assay kits for detection, thereby reducing application costs.

We observed that HRD was significantly positively correlated with the CIMPTGV predicted risk score. Homologous recombination is a mechanism for repairing double-stranded DNA breaks. However, this mechanism is often deficient in breast and ovarian cancer patients, making them more sensitive to Poly (ADP-ribose) polymerase inhibitor therapy.[47,48] It is increasingly recognized that HRD underlies the development of both hereditary and sporadic tumorigenesis,[49] leading to genomic instability and a cascade of downstream events.[50] In our study, we observed a co-expression between HRD and malignant tumor-related features, including the high expression and amplification of genes associated with aggressive tumor behavior, enrichment of nucleic acid metabolites, high tumor cell number, and high pathological heterogeneity. These findings highlight the intricate interplay between HRD and various cellular processes, emphasizing the importance of multimodal integration in enhancing the exploration and comprehensive understanding of the biological characteristics of breast cancer. The integration of diverse data types allows for a more nuanced analysis, contributing to a more targeted exploration of potential therapeutic targets and a holistic recognition of the underlying complexities in breast cancer biology.

Our study reveals that features from different modalities not only co-express but also exhibit complementary effects, augmenting information beyond any single modality.[20] The simplified CIMPTGV model we proposed also improves clinical applicability. By reducing model complexity, the simplified model could prevent model overfitting and guarantee the robustness of the model. It also reduces the data collection costs and increases the ratio of suitable patients, thus promoting clinical application. The multi-modality model is applicable to most clinical subgroups of patients with HR+/HER2− breast cancer and can effectively distinguish between high- and low-risk recurrence groups. For patients identified as high-risk recurrence group, intensified treatment strategies would be considered, while standard or reduced-dose treatments could be more suitable for the low-risk recurrence group. We will also design prospective clinical trials and use the simplified CIMPTGV model in our ongoing clinical trials to validate its predictive efficacy, such as the Fudan University Shanghai Cancer Center Breast Cancer Precision Platform Series study - neoadjuvant therapy (FASCINATE-N) trial (ClinicalTrials.gov identifier: NCT05582499), which is a prospective, randomized, precision-based umbrella trial.[51] Moreover, considering that multimodal data constitute a valuable clinical resource and the constraints of our patient cohort size, we opted for traditional machine learning methods instead of deep-learning approaches. In the future, we will also explore some deep-learning-based data integration methods such as attention-based weighted fusion and cross-modal mutual information maximization and compare their performance with our current models.

In conclusion, we collected data from a large-scale multi-omics cohort and developed multi-modality models for patient risk stratification. Our study demonstrated that machine learning models combining multiple modalities significantly outperformed those based on single modality. These findings underscore the substantial potential of multimodal machine learning in predicting tumor biological characteristics, encouraging further extensive investigations into multimodal integration for patient risk stratification in both the breast cancer and other cancer types.

Limitations of the study

Despite providing valuable insights, our study still has some limitations. Firstly, the single-center cohort may introduce bias, and it is necessary to evaluate the stability of our model in other external multi-omics cohorts. However, we did not find available public cohort data with complete modalities. Secondly, our cohort with complete multimodal information is still limited in size. Despite this, we have employed various approaches, including ensemble learning, regularization, and hyperparameter tuning, to prevent overfitting. Additionally, we also assessed overfitting after model development and demonstrated the stability of our CIMPTGV model. Another challenge is the difficulty of collecting complete feature information to construct the CIMPTGV model, potentially limiting its application. Our proposed simplified model could mitigate this issue to some extent. Although simplifying the model may result in some loss of orthogonal information, leading to a prediction performance that is not as strong as the full model, it still significantly distinguishes the high-risk population. The simplified model achieves an average AUC of 0.840, outperforming other commonly used dimensions in clinical practice. By selecting a few features with high feature importance scores, a corresponding kit could be designed to detect the crucial features, reducing the economic burden of implementing the simplified CIMPTGV model.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yi Xiao (yixiao11@fudan.edu.cn).

Materials availability

This study did not generate novel reagents.

Data and code availability

- The whole-exome sequencing (WES) data, CNV data, RNA sequencing data, and metabolome data for this study have been deposited into the Genome Sequence Archive (GSA) database under accession codes PRJCA017539 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017539).
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (92159301, 82341003, and 82373167), the Shanghai

Key Laboratory of Breast Cancer (12DZ2260100), the SHDC Municipal Project for Developing Emerging and Frontier Technology in Shanghai Hospitals (SHDC12021103), Shanghai Medical Innovation Research Project (22Y11912700), the Seed Program for Research and Translation of Innovative Medical Technologies of Shanghai Municipal Commission of Health (2024ZZ2023), and the Shanghai Science and Technology Innovation Action Plan (the Yangfan Project) from Shanghai Municipal Science and Technology Commission of China (23YF1406500).

AUTHOR CONTRIBUTIONS

Conceptualization, H.Z., F.Y., Y.-Z.J., Z.-M.S., and Y. Xiao; methodology, H.Z., Y. Xu, and S.Z.; formal analysis, H.Z.; investigation, H.Z., F.Y., Y. Xu, and Y. Xiao; writing – original draft, H.Z. and F.Y.; writing – review and editing, H.Z., F.Y., Y. Xu, and Y. Xiao; supervision, Y.-Z.J., Z.-M.S., and Y. Xiao. All authors approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Generation and analysis of multiomic data
  - Estimation of HRD score
  - Gene set enrichment analysis
  - Multimodal feature collection
  - Discrete variable processing
  - Feature dimension reduction
  - Feature normalization
  - Cell segmentation
  - Digital pathology feature extraction
  - Train set and test set split
  - Model training and testing
  - Feature importance
  - Time-dependent AUC score
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101924.

Received: December 27, 2023
Revised: August 11, 2024
Accepted: December 27, 2024
Published: January 22, 2025

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
2. Huppert, L.A., Gumusay, O., Idossa, D., and Rugo, H.S. (2023). Systemic therapy for hormone receptor-positive/human epidermal growth factor receptor 2-negative early stage and metastatic breast cancer. CA. Cancer J. Clin. *73*, 480–515. https://doi.org/10.3322/caac.21777.
3. Colleoni, M., Sun, Z., Price, K.N., Karlsson, P., Forbes, J.F., Thürlimann, B., Gianni, L., Castiglione, M., Gelber, R.D., Coates, A.S., and Goldhirsch, A. (2016). Annual Hazard Rates of Recurrence for Breast Cancer During 24 Years of Follow-Up: Results From the International Breast Cancer Study Group Trials I to V. J. Clin. Oncol. *34*, 927–935. https://doi.org/10.1200/jco.2015.62.3504.
4. Haricharan, S., Punturi, N., Singh, P., Holloway, K.R., Anurag, M., Schmelz, J., Schmidt, C., Lei, J.T., Suman, V., Hunt, K., et al. (2017). Loss of MutL Disrupts CHK2-Dependent Cell-Cycle Control through CDK4/6 to Promote Intrinsic Endocrine Therapy Resistance in Primary Breast Cancer. Cancer Discov. *7*, 1168–1183. https://doi.org/10.1158/2159-8290.Cd-16-1179.
5. Ellis, M.J., Suman, V.J., Hoog, J., Lin, L., Snider, J., Prat, A., Parker, J.S., Luo, J., DeSchryver, K., Allred, D.C., et al. (2011). Randomized phase II neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: clinical and biomarker outcomes and predictive value of the baseline PAM50-based intrinsic subtype–ACOSOG Z1031. J. Clin. Oncol. *29*, 2342–2349. https://doi.org/10.1200/jco.2010.31.6950.
6. Ellis, M.J., Suman, V.J., Hoog, J., Goncalves, R., Sanati, S., Creighton, C.J., DeSchryver, K., Crouch, E., Brink, A., Watson, M., et al. (2017). Ki67 Proliferation Index as a Tool for Chemotherapy Decisions During and After Neoadjuvant Aromatase Inhibitor Treatment of Breast Cancer: Results From the American College of Surgeons Oncology Group Z1031 Trial (Alliance). J. Clin. Oncol. *35*, 1061–1069. https://doi.org/10.1200/jco.2016.69.4406.
7. Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., Peto, R., Pritchard, K.I., Bergh, J., Dowsett, M., et al. (2017). 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. N. Engl. J. Med. *377*, 1836–1846. https://doi.org/10.1056/NEJMoa1701830.
8. Hanker, A.B., Sudhan, D.R., and Arteaga, C.L. (2020). Overcoming Endocrine Resistance in Breast Cancer. Cancer Cell *37*, 496–513. https://doi.org/10.1016/j.ccell.2020.03.009.
9. Castellano, I., Chiusa, L., Vandone, A.M., Beatrice, S., Goia, M., Donadio, M., Arisio, R., Muscarà, F., Durando, A., Viale, G., et al. (2013). A simple and reproducible prognostic index in luminal ER-positive breast cancers. Ann. Oncol. *24*, 2292–2297. https://doi.org/10.1093/annonc/mdt183.
10. Constantinidou, A., Marcou, Y., Toss, M.S., Simmons, T., Bernhisel, R., Hughes, E., Probst, B., Meek, S., Kakouri, E., Georgiou, G., et al. (2022). Clinical Validation of EndoPredict in Pre-Menopausal Women with ER-Positive, HER2-Negative Primary Breast Cancer. Clin. Cancer Res. *28*, 4435–4443. https://doi.org/10.1158/1078-0432.Ccr-22-0619.
11. Miller, P.C., Clarke, J., Koru-Sengul, T., Brinkman, J., and El-Ashry, D. (2015). A novel MAPK-microRNA signature is predictive of hormone-therapy resistance and poor outcome in ER-positive breast cancer. Clin. Cancer Res. *21*, 373–385. https://doi.org/10.1158/1078-0432.Ccr-14-2053.
12. Kurland, B.F., Peterson, L.M., Lee, J.H., Schubert, E.K., Currin, E.R., Link, J.M., Krohn, K.A., Mankoff, D.A., and Linden, H.M. (2017). Estrogen Receptor Binding (18F-FES PET) and Glycolytic Activity (18F-FDG PET) Predict Progression-Free Survival on Endocrine Therapy in Patients with ER+ Breast Cancer. Clin. Cancer Res. *23*, 407–415. https://doi.org/10.1158/1078-0432.Ccr-16-0362.
13. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med. *351*, 2817–2826. https://doi.org/10.1056/NEJMoa041588.
14. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536. https://doi.org/10.1038/415530a.
15. Prat, A., Parker, J.S., Fan, C., Cheang, M.C.U., Miller, L.D., Bergh, J., Chia, S.K.L., Bernard, P.S., Nielsen, T.O., Ellis, M.J., et al. (2012). Concordance among gene expression-based predictors for ER-positive breast cancer

treated with adjuvant tamoxifen. Ann. Oncol. *23*, 2866–2873. https://doi.org/10.1093/annonc/mds080.

16. Gui, C.P., Chen, Y.H., Zhao, H.W., Cao, J.Z., Liu, T.J., Xiong, S.W., Yu, Y.F., Liao, B., Cao, Y., Li, J.Y., et al. (2023). Multimodal recurrence scoring system for prediction of clear cell renal cell carcinoma outcome: a discovery and validation study. Lancet. Digit. Health *5*, e515–e524. https://doi.org/10.1016/s2589-7500(23)00095-x.
17. Chen, R.J., Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., and Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell *40*, 865–878.e6. https://doi.org/10.1016/j.ccell.2022.07.004.
18. Boehm, K.M., Aherne, E.A., Ellenson, L., Nikolovski, I., Alghamdi, M., Vázquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., et al. (2022). Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nat. Cancer *3*, 723–733. https://doi.org/10.1038/s43018-022-00388-9.
19. Sammut, S.J., Crispin-Ortuzar, M., Chin, S.F., Provenzano, E., Bardwell, H.A., Ma, W., Cope, W., Dariush, A., Dawson, S.J., Abraham, J.E., et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. Nature *601*, 623–629. https://doi.org/10.1038/s41586-021-04278-5.
20. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., and Shah, S.P. (2022). Harnessing multimodal data integration to advance precision oncology. Nat. Rev. Cancer *22*, 114–126. https://doi.org/10.1038/s41568-021-00408-3.
21. Acosta, J.N., Falcone, G.J., Rajpurkar, P., and Topol, E.J. (2022). Multimodal biomedical AI. Nat. Med. *28*, 1773–1784. https://doi.org/10.1038/s41591-022-01981-2.
22. Zhang, Y., Zhou, Y., Mao, F., Yao, R., and Sun, Q. (2020). Ki-67 index, progesterone receptor expression, histologic grade and tumor size in predicting breast cancer recurrence risk: A consecutive cohort study. Cancer Commun. *40*, 181–193. https://doi.org/10.1002/cac2.12024.
23. DeCensi, A., Guerrieri-Gonzaga, A., Gandini, S., Serrano, D., Cazzaniga, M., Mora, S., Johansson, H., Lien, E.A., Pruneri, G., Viale, G., and Bonanni, B. (2011). Prognostic significance of Ki-67 labeling index after short-term presurgical tamoxifen in women with ER-positive breast cancer. Ann. Oncol. *22*, 582–587. https://doi.org/10.1093/annonc/mdq427.
24. Sparano, J.A., and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. J. Clin. Oncol. *26*, 721–728. https://doi.org/10.1200/jco.2007.15.1068.
25. Moynahan, M.E., and Jasin, M. (2010). Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nat. Rev. Mol. Cell Biol. *11*, 196–207. https://doi.org/10.1038/nrm2851.
26. Hu, L., Sham, J.S.T., Xie, D., Wen, J.M., Wang, W.S., Wang, Y., and Guan, X.Y. (2007). Up-regulation of fibroblast growth factor 3 is associated with tumor metastasis and recurrence in human hepatocellular carcinoma. Cancer Lett. *252*, 36–42. https://doi.org/10.1016/j.canlet.2006.12.003.
27. Tai, A.L.S., Sham, J.S.T., Xie, D., Fang, Y., Wu, Y.L., Hu, L., Deng, W., Tsao, G.S.W., Qiao, G.B., Cheung, A.L.M., and Guan, X.Y. (2006). Co-overexpression of fibroblast growth factor 3 and epidermal growth factor receptor is correlated with the development of nonsmall cell lung carcinoma. Cancer *106*, 146–155. https://doi.org/10.1002/cncr.21581.
28. Ngan, E.S.W., Ma, Z.Q., Chua, S.S., DeMayo, F.J., and Tsai, S.Y. (2002). Inducible expression of FGF-3 in mouse mammary gland. Proc. Natl. Acad. Sci. USA *99*, 11187–11192. https://doi.org/10.1073/pnas.172366199.
29. Morini, M., Astigiano, S., Mora, M., Ricotta, C., Ferrari, N., Mantero, S., Levi, G., Rossini, M., and Barbieri, O. (2000). Hyperplasia and impaired involution in the mammary gland of transgenic mice expressing human FGF4. Oncogene *19*, 6007–6014. https://doi.org/10.1038/sj.onc.1204011.
30. Luo, M.L., Shen, X.M., Zhang, Y., Wei, F., Xu, X., Cai, Y., Zhang, X., Sun, Y.T., Zhan, Q.M., Wu, M., and Wang, M.R. (2006). Amplification and overexpression of CTTN (EMS1) contribute to the metastasis of esophageal squamous cell carcinoma by promoting cell migration and anoikis resistance. Cancer Res. *66*, 11690–11699. https://doi.org/10.1158/0008-5472.Can-06-1484.
31. Rothschild, B.L., Shim, A.H., Ammer, A.G., Kelley, L.C., Irby, K.B., Head, J.A., Chen, L., Varella-Garcia, M., Sacks, P.G., Frederick, B., et al. (2006). Cortactin overexpression regulates actin-related protein 2/3 complex activity, motility, and invasion in carcinomas with chromosome 11q13 amplification. Cancer Res. *66*, 8017–8025. https://doi.org/10.1158/0008-5472.Can-05-4490.
32. Zhao, S., Chen, D.P., Fu, T., Yang, J.C., Ma, D., Zhu, X.Z., Wang, X.X., Jiao, Y.P., Jin, X., Xiao, Y., et al. (2023). Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. Nat. Commun. *14*, 6796. https://doi.org/10.1038/s41467-023-42504-y.
33. Regan, M.M., Neven, P., Giobbie-Hurder, A., Goldhirsch, A., Ejlertsen, B., Mauriac, L., Forbes, J.F., Smith, I., Láng, I., Wardley, A., et al. (2011). Assessment of letrozole and tamoxifen alone and in sequence for postmenopausal women with steroid hormone receptor-positive breast cancer: the BIG 1-98 randomised clinical trial at 8·1 years median follow-up. Lancet Oncol. *12*, 1101–1108. https://doi.org/10.1016/s1470-2045(11)70270-4.
34. Filipits, M., Rudas, M., Jakesz, R., Dubsky, P., Fitzal, F., Singer, C.F., Dietze, O., Greil, R., Jelen, A., Sevelda, P., et al. (2011). A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin. Cancer Res. *17*, 6012–6020. https://doi.org/10.1158/1078-0432.Ccr-11-0926.
35. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. *98*, 262–272. https://doi.org/10.1093/jnci/djj052.
36. Wang, Y., Klijn, J.G.M., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet *365*, 671–679. https://doi.org/10.1016/s0140-6736(05)17947-1.
37. Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167. https://doi.org/10.1200/jco.2008.18.1370.
38. Shoshan, Y., Bakalo, R., Gilboa-Solomon, F., Ratner, V., Barkan, E., Ozery-Flato, M., Amit, M., Khapun, D., Ambinder, E.B., Oluyemi, E.T., et al. (2022). Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis. Radiology *303*, 69–77. https://doi.org/10.1148/radiol.211105.
39. Lu, M.T., Raghu, V.K., Mayrhofer, T., Aerts, H.J.W.L., and Hoffmann, U. (2020). Deep Learning Using Chest Radiographs to Identify High-Risk Smokers for Lung Cancer Screening Computed Tomography: Development and Validation of a Prediction Model. Ann. Intern. Med. *173*, 704–713. https://doi.org/10.7326/m20-1868.
40. Huang, P., Lin, C.T., Li, Y., Tammemagi, M.C., Brock, M.V., Atkar-Khattra, S., Xu, Y., Hu, P., Mayo, J.R., Schmidt, H., et al. (2019). Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. Lancet. Digit. Health *1*, e353–e362. https://doi.org/10.1016/s2589-7500(19)30159-1.
41. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118. https://doi.org/10.1038/nature21056.
42. Yi, Z., Hu, S., Lin, X., Zou, Q., Zou, M., Zhang, Z., Xu, L., Jiang, N., and Zhang, Y. (2022). Machine learning-based prediction of invisible intraprostatic prostate cancer lesions on (68) Ga-PSMA-11 PET/CT in patients with primary prostate cancer. Eur. J. Nucl. Med. Mol. Imaging *49*, 1523–1534. https://doi.org/10.1007/s00259-021-05631-6.

43. Chang, X., Wang, J., Zhang, G., Yang, M., Xi, Y., Xi, C., Chen, G., Nie, X., Meng, B., and Quan, X. (2023). Predicting colorectal cancer microsatellite instability with a self-attention-enabled convolutional neural network. Cell Rep. Med. *4*, 100914. https://doi.org/10.1016/j.xcrm.2022.100914.

44. Foersch, S., Eckstein, M., Wagner, D.C., Gach, F., Woerl, A.C., Geiger, J., Glasner, C., Schelbert, S., Schulz, S., Porubsky, S., et al. (2021). Deep learning for diagnosis and survival prediction in soft tissue sarcoma. Ann. Oncol. *32*, 1178–1187. https://doi.org/10.1016/j.annonc.2021.06.007.

45. Vanguri, R.S., Luo, J., Aukerman, A.T., Egger, J.V., Fong, C.J., Horvat, N., Pagano, A., Araujo-Filho, J.d.A.B., Geneslaw, L., Rizvi, H., et al. (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. Nat. Cancer *3*, 1151–1164. https://doi.org/10.1038/s43018-022-00416-8.

46. Jin, X., Zhou, Y.F., Ma, D., Zhao, S., Lin, C.J., Xiao, Y., Fu, T., Liu, C.L., Chen, Y.Y., Xiao, W.X., et al. (2023). Molecular classification of hormone receptor-positive HER2-negative breast cancer. Nat. Genet. *55*, 1696–1708. https://doi.org/10.1038/s41588-023-01507-7.

47. Mirza, M.R., Monk, B.J., Herrstedt, J., Oza, A.M., Mahner, S., Redondo, A., Fabbro, M., Ledermann, J.A., Lorusso, D., Vergote, I., et al. (2016). Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. N. Engl. J. Med. *375*, 2154–2164. https://doi.org/10.1056/NEJMoa1611310.

48. Robson, M., Im, S.A., Senkus, E., Xu, B., Domchek, S.M., Masuda, N., Delaloge, S., Li, W., Tung, N., Armstrong, A., et al. (2017). Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. N. Engl. J. Med. *377*, 523–533. https://doi.org/10.1056/NEJMoa1706450.

49. Evers, B., Helleday, T., and Jonkers, J. (2010). Targeting homologous recombination repair defects in cancer. Trends Pharmacol. Sci. *31*, 372–380. https://doi.org/10.1016/j.tips.2010.06.001.

50. Peng, G., Chun-Jen Lin, C., Mo, W., Dai, H., Park, Y.Y., Kim, S.M., Peng, Y., Mo, Q., Siwko, S., Hu, R., et al. (2014). Genome-wide transcriptome profiling of homologous recombination DNA repair. Nat. Commun. *5*, 3361. https://doi.org/10.1038/ncomms4361.

51. Zuo, W.J., Chen, L., Shen, Y., Wang, Z.H., Liu, G.Y., Yu, K.D., Di, G.H., Wu, J., Li, J.J., and Shao, Z.M. (2024). Rational and trial design of FASCINATE-N: a prospective, randomized, precision-based umbrella trial. Ther. Adv. Med. Oncol. *16*, 17588359231225032. https://doi.org/10.1177/17588359231225032.

52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research *12*, 2825–2830.

53. Polsterl, S. (2020). scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. J. Mach. Learn. Res. *21*, 1–6.

54. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

55. McKinney, W. (2010). Data structures for statistical computing in python. https://doi.org/10.25080/Majora-92bf1922-00a.

56. Hammond, M.E.H., Hayes, D.F., Dowsett, M., Allred, D.C., Hagerty, K.L., Badve, S., Fitzgibbons, P.L., Francis, G., Goldstein, N.S., Hayes, M., et al. (2010). American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). Arch. Pathol. Lab Med. *134*, e48–e72. https://doi.org/10.5858/134.7.e48.

57. Jiang, Y.Z., Ma, D., Jin, X., Xiao, Y., Yu, Y., Shi, J., Zhou, Y.F., Fu, T., Lin, C.J., Dai, L.J., et al. (2024). Integrated multiomic profiling of breast cancer in the Chinese population reveals patient stratification and therapeutic vulnerabilities. Nat. Cancer *5*, 673–690. https://doi.org/10.1038/s43018-024-00725-0.

58. Jiang, Y.Z., Ma, D., Jin, X., Xiao, Y., Yu, Y., Shi, J., Zhou, Y.F., Fu, T., Lin, C.J., Dai, L.J., et al. (2023). Integrated molecular portraits of breast cancer. Nat. Protoc. https://doi.org/10.21203/rs.3.pex-2435/v1.

59. Jiang, Y.Z., Ma, D., Suo, C., Shi, J., Xue, M., Hu, X., Xiao, Y., Yu, K.D., Liu, Y.R., Yu, Y., et al. (2019). Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. Cancer Cell *35*, 428–440.e5. https://doi.org/10.1016/j.ccell.2019.02.001.

60. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf. *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

61. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human WES data	This paper	https://ngdc.cncb.ac.cn/bioproject/browse; ID: PRJCA017539
Human CNV data	This paper	https://ngdc.cncb.ac.cn/bioproject/browse; ID: PRJCA017539
Human RNA-seq data	This paper	https://ngdc.cncb.ac.cn/bioproject/browse; ID: PRJCA017539
Human metabolome data	This paper	https://ngdc.cncb.ac.cn/bioproject/browse; ID: PRJCA017539
Software and algorithms		
R	R Core Team	https://www.r-project.org/
Python (v.3.11)	Python software Foundation. Python Language Reference, version 3.11.	https://www.python.org
Biorender	N/A	https://biorender.com/
Scikit-learn (v.1.2.2)	Virtanen et al.[52]	https://github.com/scipy/scipy
Scikit-survival (v.0.22.1)	Pölsterl[53]	https://scikit-survival.readthedocs.io/en/stable/index.html
Numpy (v.1.24.3)	Harris et al.[54]	https://github.com/numpy/numpy
Pandas (v.1.5.3)	McKinney[55]; Pandas Development Team	https://github.com/pandas-dev/pandas

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The patients (n = 579, 23–90 years of age) analyzed in this study were from a retrospective selection of female patients who had been diagnosed with malignant breast cancer in our center which has been published before.[46] Our cohort consisted of pre-treatment Chinese female patients from January 2013 to December 2014 without intentional selection (n = 478). We also selected Chinese female patients who experienced relapse after surgery between January 2009 and December 2016 (n = 101). There was no influence of sex or gender on the results of our study. Patients included had to meet the following criteria: (1) female patients diagnosed with HR+ (ER + or PR+) and HER2-unilateral invasive breast carcinoma; (2) pathological examination of tumor samples was conducted by the pathology department of FUSCC, with the ER, PR, and HER2 status verified by two experienced pathologists using immunochemical analysis and *in situ* hybridization. ER/PR negativity in IHC testing was defined as less than 1% positive staining of cells, following the guidelines set forth by the American Society of Clinical Oncology/College of American Pathologists.[56] (3) There was no sign of distant metastasis and no pre-treatment at the time of diagnosis. (4) Frozen samples were supplied for further analysis. Patients with inflammatory breast cancer or breast carcinoma *in situ* were excluded from the study. The studies were conducted in accordance with the Declaration of Helsinki. All analyses were approved by the independent ethics committee/institutional review board of Fudan University Shanghai Cancer Center. All participants in the study provided written informed consent before inclusion and there was no compensation during the study.

Follow-up of patients in this cohort was completed on June 30, 2021, and the median length of follow-up was 79.1 months (IQR = 72.1–93.1 months). Relapse-free survival (RFS) was defined as the time from diagnosis to first recurrence, contralateral breast cancer or death from any cause. Overall survival (OS) was defined as the length of time from the date of surgery to death. Distant metastasis-free survival (DMFS) was defined as the time from the date of surgery to the first detection of distant metastasis or death.

METHOD DETAILS

Generation and analysis of multiomic data

The detailed methods for generation and analysis of genomic, transcriptomic and metabolomic data are described in a previously published article,[57] which has been deposited on the Nature Portfolio Protocol Exchange platform.[58]

Estimation of HRD score

The HRD score was calculated as the unweighted sum of the heterozygosity (LOH) score, telomeric allelic imbalance (TAI) score, and modified large-scale state transition (LSTm) score. The LOH score represents the number of sub-chromosomal regions with loss of LOH that are longer than 15 Mb. The LST score counts the number of chromosomal breaks of at least 10 Mb between adjacent regions after filtering out regions shorter than 3 Mb. To account for the effect of ploidy, the LST score was adjusted using the following formula: LSTm = LST −15.5P (where P represents ploidy). The TAI score was defined as the number of regions of allelic imbalance

that extend to one of the subtelomeres without crossing the centromere. The detailed calculation of these scores has been previously described.[59]

Gene set enrichment analysis

Single-sample gene set enrichment analysis (ssGSEA) was performed using the GSVA R package (version 1.48.3) to generate pathway scores based on transcriptomic signatures.[60]

Multimodal feature collection

Commonly used clinical characteristics such as age, BMI, vessel invasion, menopause, tumor stage and lymph node stage were input as C (6 features). Immunohistochemistry (IHC) features such as ER percent, PR percent and Ki-67 proliferation index were input as I (3 features). The ssGSEA scores of hallmark gene sets, Kyoto Encyclopedia of Genes and Genome (KEGG) gene sets and immune cell type fractions inferred by using CIBERSORT[61] were input as transcriptomic features namely T (190 features). Metabolomic features, namely M, consisted of polar metabolites and lipids detected by liquid chromatography with tandem mass spectrometry (LC-MS/MS) (1981 features). The mutation status of genes with the top 40 mutation frequencies, tumor mutation burden (TMB) score, loss of heterozygosity (LOH) score, and homologous recombination deficiency (HRD) score were input as genomic features as G (43 features). Gistic peak values were input as CNV features as V (76 features). Pathological features including nuclear features, topological features and composition features were input as P (59 features).

Discrete variable processing

We conducted separate preprocessing for each modality. Within the clinical dimension, which includes discrete variables such as clinical stage, we utilized the OneHotEncoder function in scikit-learn package to incorporate these crucial discrete features into our model.

Feature dimension reduction

Feature dimension reduction is performed separately in different modalities. First, in each modality, we calculated the correlation values between features. Features pairs with correlations higher than 0.9 were selected. Then we retained the feature with a higher score by using CoxPHSurvivalAnalysis in scikit-survival and dropped the feature with a lower score. We performed the above steps and obtained the filtered feature set in each modality. Second, the filtered feature set in each modality was further filtered using selectkbest in scikit-learn. To balance the features across different dimensions and reduce model complexity and training time, we included all features from clinical and IHC modalities, and selected the top 15 features with the largest univariate Cox hazard score from other modalities.

Feature normalization

The feature set after dimension reduction is normalized using the standardscaler function in the scikit-learn package before model training or testing.

Cell segmentation

We used the Hover-Net to locate nuclei on whole slide images (WSIs) and predict their cell types. We utilized a pre-trained model on the publicly available PanNuke dataset, capable of recognizing five distinct nuclear types. These types include tumor, inflammatory, stromal, normal and dead nuclei. For each WSI, Hover-Net output the information on centroid, contour and nuclear type of detected nuclei for the following feature extraction.

Digital pathology feature extraction

Nuclear morphological features and texture features were extracted for individual cells. Morphological features describe nuclear area, perimeter, elongation and curvature. Texture features characterize features on the basis of gray-level co-occurrence matrix and statistics of pixel intensity within the nuclear contour. After this, a series of statistics (mean, median, standard variation, skewness and kurtosis) were calculated to integrate the cell-level features into WSI-level features. In addition to nuclear morphological ad texture features, we computed tissue composition features for each WSI. We tessellated all WSIs into 256 × 256 pixel tiles at a 20×magnification. Then, our established classifier was applied to categorize all image tiles, assigning each tile to one of the five tissue types. The percentage of image tiles of each tissue type was calculated, forming the tissue composition features for the WSI.

Topological features were extracted to characterize the intercellular relationship at the single-cell resolution through a graph-based method proposed by us.[32] For each pair of cell types, a graph was constructed denoted as Graph_C1-C2, where C1 and C2 denoted the cell types. We focused on tumor, inflammatory and stroma cells. For the construction of each graph, k-nearest neighbor algorithm was used for edge configuration. Each cell is connected to its five nearest cell according to the Euclidean distance and the edges longer than a threshold of 100 pixel (25 μm) were deleted. For each cell, the topological features were computed separately for each graph it appeared and were concatenated to constitute its final comprehensive topological features.

The nuclear features and topological features for patients were obtained by averaging the topological features of each cell type. Detailed descriptions of all morphological, texture and topological features were presented in Table S3.

Train set and test set split

The train set and test set were divided in a 4:1 ratio in each model cohort by using the random stratified sampling method, with relapse status used as the stratification factor.

Model training and testing

The machine learning framework was built on Python (version 3.11), scikit-learn (version 1.2.2), scikit-survival (version 0.22.1), numpy (version 1.24.3) and pandas (version 1.5.3). After feature selection and dimension reduction, *Z* score scaling was applied to the remaining features. Then the scaled feature matrix was input into five algorithms, including Cox's proportional hazard's model, survival support vector machine, random survival forest model, DeepSurv non-linear model and gradient boosting survival model. All hyperparameters were optimized using a randomized 1,000-step 5-fold cross-validation search to maximize the average cross-validation C-index and the model with the highest average cross-validation C-index was selected as the final model. The performance of each model was assessed by the C-index with 95% CI in the test set by using 1000-fold bootstrapping. Detailed descriptions of the hyperparameters that were modified from their default values, and the range of hyperparameters for selecting were presented in Table S2 and Table S4.

Feature importance

Feature importance was determined for each feature in the final best model, namely the CIMPTGV model. We dropped each feature in the model and used the reduced C-index value as the importance score for the feature. Higher feature importance scores mean more important features in the model.

Time-dependent AUC score

The time-dependent AUC score was calculated by using the cumulative dynamic AUC function in scikit-survival (version 0.22.1).[53] Cumulative cases are patients who experienced relapse prior to or at time t (ti $\leq$ t), whereas dynamic controls are patients with ti > t. By using the risk score $\widehat{f}(\mathbf{x}_i)$ of the *i*-th case, the cumulative AUC at time *t* is defined as:

$$\widehat{\mathrm{AUC}}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(y_j > t) I(y_i \leq t)\omega_i I(\widehat{f}(\mathbf{x}_j) \leq \widehat{f}(\mathbf{x}_i))}{\left(\sum_{i=1}^{n} I(y_i > t)\right)\left(\sum_{i=1}^{n} I(y_i \leq t)\omega_i\right)}$$

ω_i are inverse probability of censoring weights (IPCW) estimated by accessing survival times in the training data.

QUANTIFICATION AND STATISTICAL ANALYSIS

The Mann-Whitney *U*-test was utilized to compare continuous variables and ordered categorical variables. The 95% confidence intervals of the c-index of the test set were determined by using 1000-fold bootstrapping. The Kaplan-Meier product-limit method of survival and log rank test were used to compare and explore differences in RFS, OS and DMFS between high-risk and low-risk groups. A Cox proportional hazard regression model was performed to generate hazard ratios (HRs) and 95% confidence intervals. All tests were two-sided. Analyses were performed using the R software (version 4.3.1) and Python (version 3.11). Statistical details (including the statistical tests used, definition of center and spread, and statistical significance degrees) can be found in the corresponding figures and figure legends.