Breast-NEOprAIdict: a deep learning solution for predicting pathological complete response on biopsies of breast cancer patients treated with neoadjuvant chemotherapy.

Natalia Fernanda Valderrama [1]*, Louis-Oscar Morel [1]*, Daniel Tshokola Mweze [1], Valentin Derangère [2,3,4], Isabelle Desmoulins [5], Didier Mayeur [5], Courèche Kaderbhai [5], Silvia Ilie [5], Audrey Hennequin [5], Nicolas Roussot [5], Antony Bergeron [2], Françoise Beltjens [2], Carlo Pescia[6], Henri-Philippe Morel [7], Charles Coutant [8], Laurent Arnould [2], Nathan Vinçon [1‡], Sylvain Ladoire [2,3,4,5,9 ‡]

1. Ummon HealthTech SAS, Dijon, France
2. Department of Biology and Pathology of tumors, Georges Francois Leclerc Cancer Centre, Dijon, France
3. Platform of Transfer in Biological Oncology, Georges François Leclerc Cancer Center – Dijon – France
4. University of Burgundy-Franche Comté, 21000 Dijon, France
5. Department of Medical Oncology, Georges Francois Leclerc Cancer Centre, Dijon, France
6. Department of Molecular Medicine, University of Milan
7. Technipath, Lyon, France
8. Department of Surgical Oncology, Georges Francois Leclerc Cancer Centre, Dijon, France
9. INSERM U1231, 21000 Dijon, France

Keywords: Breast cancer; Neoadjuvant chemotherapy; deep learning; artificial intelligence; pathology; stratification; pathological complete response

*Natalia Fernanda Valderrama and Louis-Oscar Morel contributed equally to this work.

‡ Sylvain Ladoire and Nathan Vinçon contributed equally to this work.

‡ corresponding authors

Funding: The PRIMUNEO cohort analyzed in this report was selected by the PHRC-Cancer call for proposals and funded through a grant from the French Ministry of Health PHRC-K2011.

Statement of translational relevance

This study introduces Breast-NEOprAIdict, a novel deep learning model designed to predict pathological complete response (pCR) in early breast cancer patients undergoing standard neoadjuvant chemotherapy. Leveraging standard histological images stained with hematoxylin and eosin (H&E), this model is easily integrable into existing diagnostic workflows. Breast-NEOprAIdict offers a direct prediction of tumor chemosensitivity, providing insights that are more closely aligned with pCR outcomes than current classification methods based on staging and grading. The adoption of this tool in clinical practice could enhance personalized treatment strategies, enabling more accurate identification of patients who are likely to benefit from standard chemotherapy and improving overall patient management in breast cancer care.

Abstract

In precision medicine, the prediction of tumor chemosensitivity is of major importance to offer cancer patients the best possible treatment from the outset. In this study, we introduce Breast-NEOprAIdict, a deep learning model designed to predict the occurence of pathological complete response (pCR) in early breast cancer (eBC) patients treated with standard neoadjuvant chemotherapy (NAC).

This prediction is based on an analysis of the initial tumor diagnostic biopsy. To this end, we used two extensive cohorts (total n=1140 patients) spanning various molecular subtypes of eBC (HER2-amplified (HER2+), estrogen-receptor positive/HER2 non amplified (ER+/HER2-), and triple-negative (TN) tumors): the PRIMUNEO prospective cohort (n=500) for training and internal validation and the CGFL Breast Cancer Neoadjuvant database (n=640) for external validation.

Breast-NEOprAIdict demonstrated good performance on the external validation dataset for HER2+ tumors (Area Under the Curve (AUC): 0.652 (P = 0.001), Odds Ratio (OR): 2.42 (P = 0.0131)), ER+/HER2- tumors (AUC: 0.814 (P = 0.003), OR: 20.56 (P = 0.00413)) and TN tumors (AUC: 0.677 (P = 0.001), OR: 3.44 (P=0.00373)) compared to standard clinicopathological features. We also evaluated the robustness of our algorithm through testing on several scanned sections per patient. Breast-NEOprAIdict exhibited strong consistency in the external validation cohort, with a Pearson correlation coefficient of 0.933 (P <0.001) for HER2+, 0.932 (P <0.001) for ER+/HER2- tumors, and 0.939 (P <0.001) for TN.

Breast-NEOprAIdict is a new tool for identifying eBC that are differentially sensitive to standard NAC, and could help to select the most appropriate treatment strategy in HER2+, ER+/HER2- and TN eBC.

Introduction

Breast cancer (BC) is the leading cause of cancer-related deaths for women in Europe and worldwide[1]. Adjuvant or neoadjuvant systemic chemotherapy is a standard treatment for high-risk early breast cancer (eBC) and consists of performing chemotherapy after or before surgery, respectively. Chemotherapy is recommended for the vast majority of patients with HER2-amplified (HER2+) or triple-negative (TN) eBC, as well as for patients with ER+/HER2- tumors who have the most significant risk of metastatic relapse. The decision to administer systemic chemotherapy, either in an adjuvant or neoadjuvant setting, is made by the oncologist based on each individual patient's risk factors for relapse[1]. Neoadjuvant chemotherapy (NAC)[2] can be proposed to patients to reduce the size of the primary tumor before surgery, to facilitate conservative surgery, but also to administer systemic treatment very early on, with a view to treating any micrometastatic disease as quickly as possible. In the Early Breast Cancer Trialists' Collaborative Group (EBCTCG)[3] meta-analysis, there was no significant difference between patients treated with a neoadjuvant scheme in terms of distant recurrence, breast cancer mortality, or death from any other cause, compared to the same standard adjuvant chemotherapies. An advantage of NAC is the ability to evaluate the chemosensitivity of each patient's tumor to the standard chemotherapy on the surgical specimen. Thus, patients with complete pathological response (pCR, defined as no residual invasive cancer cells either in the breast or axillary lymph nodes: pT0 and pN0[4]) after NAC have significantly better relapse-free, and overall survival compared to patients with residual disease (RD) on the surgical specimen[4,5] (especially for HER2-amplified and TN breast cancer). For these two BC subtypes, patients with RD can be selected after surgery for post operative, adjuvant treatment intensification[6,7]. For the ER+/HER2- eBC subtype, achieving pCR is significantly less common and has a reduced prognostic value[4,5]. Nevertheless, there is a small subset of chemosensitive ER+/HER2- tumors that achieve pCR, and for which chemotherapy is probably of great interest in the systemic treatment plan. Unfortunately, there is currently no effective predictive marker for identifying these tumors prior to treatment.

Therefore, predicting tumor chemosensitivity (in terms of pCR) to conventional chemotherapy regimens right from the initial diagnosis would be highly beneficial. Such prediction would enable medical oncologists and surgeons to identify patients who are likely to achieve pCR with standard treatment, and who may therefore not require intensification of systemic therapy. Conversely, they would also be able to identify those with tumors unlikely to achieve pCR after standard treatment, indicating a potential need for treatment intensification early on, or for consideration of alternative therapies. In this context, our study analyzed two extensive French cohorts of eBC patients undergoing standard NAC to develop and validate a deep learning model that uses whole slide images (WSI) from initial tumor biopsies to predict the likelihood of pCR following NAC.

Patients and Methods

Datasets:

The data analyzed in this study is sourced from two distinct eBC patients cohorts: i) the PRIMUNEO dataset, a French multicentric prospective dataset from the PRIMUNEO study (ClinicalTrials.gov Identifier: NCT01513408), conducted between May 2012 and February 2015 at different cancer centers in France, and dedicated to the identification of predictive/prognostic histopathological factors in eBC patients treated with standard NAC (this study was funded by a grant from the French Ministry of Health PHRC-K2011); and ii) the CGFL breast cancer neoadjuvant dataset, a retrospective single-center database generated from the neoadjuvant treated population in a single French cancer center (Centre Georges François Leclerc, Dijon) between the early 2000s and 2022.

- *PRIMUNEO Database:*

A total of 500 patients who received NAC between May 2012 and February 2015 in 12 different cancer centers in France were enrolled in the PRIMUNEO study. Briefly, the main inclusion criteria were: female patients between the ages of 18 and 80 years, with proven localized breast cancer, regardless of the histological type or molecular subtype (HER2-amplified, ER+/HER2-, TNBC); treated with standard NAC incorporating taxanes ± anthracyclines (treatment protocol at the physician's discretion, see **Table 1** for more information). The main exclusion criteria were: metastatic breast cancer; neoadjuvant radiotherapy; patient not amenable to surgery; and ongoing therapy for any other type of cancer.

A total of 438 hematoxylin and eosin (H&E)-stained slides were used from this dataset; 62 patients were excluded because they met the original PRIMUNEO study exclusion criteria (n=17) or had no available molecular subtype status (n=15), no available pathological response report (n=31) or no information of the laboratory origin (n=4) (**Supplementary Figure S1A**). Some patients had more than one exclusion criteria.

A more detailed description of this cohort is provided in **Table 1** and **Supplementary Table S1**.

- *CGFL Breast Cancer Neoadjuvant dataset:*

The CGFL breast cancer neoadjuvant dataset comprises 1319 digitized WSI of initial diagnostic tumor biopsies obtained from 640 patients who received NAC between January 2000 and January 2022. Of these, 150 were excluded. Exclusion was for the following reasons: poor quality WSI (n=5), no molecular subtype status (n=51) or no available pathological response report (n=110) (**Supplementary Figure S1B**). Some patients had more than one exclusion criteria.

A more detailed description of this cohort is provided in **Table 1** and **Supplementary Table S2**.

Pathological evaluation:

All tissue sections, initial biopsies, and surgical specimens after NAC were examined microscopically by experienced pathologists (LA, AB, FB). For each patient included in the study, a tumor block from the initial biopsy and a representative block of residual tumor was chosen by the reference pathologist in each investigating center. Depending on the quality of pathological response, this representative block could come from an area of complete tumoral regression (in case of pCR), an area of partial tumor regression, or an area of unmodified residual tumor (in case of non-pCR). Pathological complete response (pCR) was defined as the disappearance of invasive tumor on the surgical specimen and in the lymph nodes after NAC (pT0 pN0)[4]. The pathological response was dichotomized as pCR *vs* residual disease (RD). In a subsequent analysis, RD was subdivided into two groups: "no response" and "partial response." "No response" was defined as cases where the AJCC stage either remained the same or progressed compared to the stage at diagnosis. "Partial response" referred to cases where the AJCC stage decreased but remained above stage 0 compared to the initial diagnosis. Tumor infiltrating lymphocyte (TIL) quantification, and Residual Cancer Burden (RCB) information were not available in these two cohorts of patients. The estrogen receptor (ER), progesterone receptor (PR), and HER2 status were assessed by immunohistochemistry (IHC). HER2 status was defined as positive only when IHC (3 +) or IHC (2+) and HER2 amplification by fluorescence in situ hybridization (FISH), while breast cancer with IHC (0/1 +) or IHC (2+) without HER2 amplification by FISH were considered as HER2-negative disease[8]. ER/PR positivity was defined as positive nucleus staining in more than 10% of tumor cells[9]. The nuclear grade was assessed based on the Nottingham grading system[10], and clinical staging according to the American Joint Committee on Cancer (AJCC) classification[11].

Internal and External validation study design:

In order to ensure that the site of origin was not biasing the prediction performance, as previously shown by Howard et al.[12], we used a stratified grouped cross-validation approach. We used 4 distinct training and test sets preserving similar pCR prevalence in each subset and without site overlap between a training set and its associated test set. **Supplementary Figure S2** displays the laboratories used in the training and test subsets, along with the number of slides with pCR and RD in each partition. Additionally, we validated our Deep Learning system in an external dataset (namely the CGFL Breast Cancer Neoadjuvant database). The training dataset used was the PRIMUNEO dataset, minus the slides coming from the CGFL database to avoid site-specific bias (these 112 patient slides were not included in the training set to avoid any information sharing between the training and the

test phase). The complete flowchart of internal and external validation sets is provided in **Supplementary Figure S1A - S1B**.

Pipeline Analysis:

- *Image processing*

The H&E-stained WSIs of initial biopsies were obtained using a Hamamatsu Nanozoomer 2.0HT scanner at 40 × magnification.

- *Pipeline*

All analyses were performed using Python 3.8. The pipeline is described in **Figure 1**.

i) Image preprocessing:

Hamamatsu NDPI files of H&E diagnostic biopsy slides from the PRIMUNEO and CGFL databases were first selected. A single slide was used for each patient for performance evaluation. We then extracted the foreground using an in-house trained U-net and tiled the images in non-overlapping patches of 600x600 pixels at a 5x resolution. The resulting patches were used as inputs of two separate neural network architectures (**Figure 1A)**. For the first architecture, we employed EfficientNetB7[13] with ImageNet-pretrained weights, adding a global average pooling layer to produce an embedding vector of size 2560 **(Figure 1B-I)**. For the second architecture, we used the Vision Transformer Small with 16 patches (ViT-S/16)[14] with weights from the Self Supervised Learning (SSL) DINO method[15] pre-trained on the TCGA dataset **(Figure 1B-II)**.

ii) Label processing:

These patches were associated with the label 1 if pCR was observed on the surgical specimen, or 0 if Residual Disease was described.

iii) Neural network training, model selection:

We developed a deep learning model based on two architectures to predict pCR from the WSI of biopsies.

Architecture 1: EfficientNet B7-Based Model. The first architecture employs the EfficientNet B7 embeddings as inputs for a multi-layer perceptron (MLP). This MLP consists of two fully connected layers with output dimensions of 64 and 16, each followed by a ReLU activation layer. Then, we used a single SoftMax layer to jointly predict the patient's pCR or RD status and molecular subtype (**Figure 1B-I**). Lastly, the slide-level prediction was calculated using the 99th percentile of the patch-level prediction values (**Figure 1C-I**). A Multiple Instance Learning (MIL) aggregation was also tried but did not provide competitive results.

Architecture 2: Vision Transformer-Based Model. The second architecture utilizes ViT-S/16 patch embeddings as inputs for an MLP, also consisting of two fully connected layers with output dimensions of 64 and 16, each followed by a ReLU activation layer. A linear layer predicts the pCR or RD status. The slide-level prediction is calculated using the 99th percentile of the patch-level prediction values (**Figure 1C-II**). A Multiple Instance Learning (MIL) aggregation was also tried but did not provide competitive results.

Models from both architectures were trained using a nested three-fold cross-validation approach. Each training set was divided into three subsets: two subsets were used for training a model, while the third subset served as the validation set. This process was repeated three times, each time using a different subset for validation, resulting in three trained models per architecture. We used the validation subset to select the training epoch with the best performance. The final slide-level pCR prediction was obtained by first averaging the predictions from the three models within each architecture. Next, we computed the harmonic mean of the ensemble predictions across both architectures **(Figure 1D and Supplementary Figure S3**).

Regarding further specifications, we trained each model over 500 000 iterations and a batch size of 32. For the EfficientNet B7-Based model we employed a cross-entropy loss and SAM wrapper[16] with an Adam optimizer[17] as the underlying optimizer and a base learning rate of 1e−3. For the Vision Transformer-Based Model, we employed a binary cross-entropy loss with an Adam optimizer and a learning rate of 1e−3.

iv) Clinical Model:

To provide a baseline for comparison with our deep learning model from WSIs, we developed a Clinical Model (CM) based on the pathological and clinical information. This model utilized the following data: age at surgery (ranging from 22 to 88 years), the time difference between biopsy and surgery, the molecular subtype (3 classes: HER2+, ER+/HER2-, and TN breast tumors (ER-/HER2-)), AJCC staging (9 classes: 0, I, IA, IB, IIA, IIB, IIIA, IIIB, IIIC), and Scarff-Bloom-Richardson (SBR) grade (3 classes: I, II, III) information. If the data was available for the patient, it was encoded in a one-hot vector; otherwise, a vector of zeros was used. The resulting 17-feature vector was used as an input for an MLP with two hidden layers of output dimensions 64 and 16. We then used the same approach as described earlier to predict pCR, with the pCR/RD status and the patient's molecular subtype combined output. To ensure a fair comparison between the models, the CM employed the same implementation details and training curriculum as detailed above.

v) Output binarization:

We binarized the predictions using different thresholds for each molecular subtype corresponding to the highest median of the pCR predictions in the validation subsets of the internal and external studies.

vi) Hardware and software specifications:

Experiments were run with a NVIDIA RTX A4000 graphic card and the following libraries: PyTorch v1.12.1, CUDA 11.5.

- *Review of High-Scoring Patch by experts*

To enhance our understanding of the results produced by Breast-NEOprAIdict, we focused on analyzing patches with the highest scores from slide predictions categorized as either the most chemosensitive or the most chemoresistant by our deep learning system, regardless of the ground truth. Specifically, we extracted the top 20 patches with the highest deep learning scores from the 20 slides with highest and lowest predictions. These selected patches were then reviewed by three independent experienced pathologists from three different laboratories (CP, LA, HPM, resp. with 5, 40 and 35 years of experience) for clustering based on shared visual characteristics.

Statistics and Metrics:

Clinical and pathological characteristics were compared between cohorts using the Chi-squared test or two-sided Fisher's exact test. Discriminatory power was measured using the area under the receiver operating characteristic curve (AUC) and statistical significance was assessed using a one-sided Mann–Whitney U test. To measure the strength of the association between the binarized predictions and actual pCR, we used the odds ratio (OR) with the Haldane-Anscombe correction[18], and a two-sided Fisher's Exact test for statistical analysis[19]. The models were systematically evaluated separately on each molecular subtype to avoid biological bias.

For the internal study, we computed the AUC for each partition, while for the OR, we concatenated the predictions of each partition to obtain a single value (**Supplementary Figure S4**). This approach was taken to mitigate the potential bias in the OR estimates due to the small sample size[20] and low prevalence rates for certain molecular subtypes (**as depicted in Supplementary Figure S3**).

Results

Study population characteristics:

Table 1 summarizes the clinical and pathological characteristics of all patients included in the study. We tested the differences in characteristics between the two datasets and found no statistically significant difference in menopausal status ($p = 0.2013$), breast surgery type ($p=0.0526$), or clinical and pathologic tumor stage (cAJCC and pAJCC), with p-values of 0.7967 and 0.2459, respectively. However, we found statistically significant differences in clinical and pathological tumor and node stage ($p < 0.001$), SBR grade ($p < 0.001$), tumor molecular subtype ($p < 0.001$), NAC treatment protocol ($p < 0.001$), and pCR status ($p = 0.0180$). Notably, the PRIMUNEO dataset exhibited a higher incidence of patients achieving pCR compared to the CGFL dataset, at 24% and 17.9% respectively. Regarding treatment, most of the patients received a combination of anthracycline and taxane therapies, 93.1% in the PRIMUNEO cohort and 57.7% in the CGFL dataset ('Other' category in **Table 1** also including intensified treatments combining both anthracyclines and taxanes). All patients with HER2 amplification were treated with tailored trastuzumab therapy along with standard chemotherapy (for a detailed description of the pCR/RD distribution across each molecular subtype, see **Supplementary Table S1**).

Following the patient selection procedure detailed in **Supplementary Figure S1**, our study cohort was refined to 928 participants from the initial 1140. This included 438 from the PRIMUNEO dataset and 490 from the CGFL Breast Cancer Neoadjuvant database. For internal validation, we used only the PRIMUNEO dataset, comprising 438 patients categorized into HER2+ (n=116, with a 37.1% pCR rate), ER+/HER2- (n=191, 10.5% pCR rate), and TN breast cancer (n=131, 32.1% pCR rate). We arranged the patients into four distinct partitions for cross-validation (**see Methods and Supplementary Figure S2**). For the external validation stage, we used the CGFL Breast Cancer Neoadjuvant database comprising 490 individuals characterized as follows: HER2+ (n=220, 21.4% pCR rate), ER+/HER2- (n=156, 3.8% pCR rate), and TN breast cancer (n=114, 30.7% pCR rate).

<u>Pathological Complete Response is predictable by a Deep Learning System using WSI on HER2+, ER+/HER2- and TN Breast initial diagnostic cancer biopsies:</u>

Internal validation study (PRIMUNEO cohort)

We introduce Breast-NEOprAIdict a deep learning model that uses WSI as input to predict pCR, The predictive capacity of Breast-NEOprAIdict (see Methods) was systematically compared it to a clinical model (CM) based on clinicopathological information (tumor molecular subtype, AJCC staging[11], and SBR grade information[10]). Internal validation was performed with cross-validation on the PRIMUNEO dataset. Comparison of the performance of the Breast-NEOprAIdict and CM in terms of AUC are shown in **Figure 2** and for OR in **Table 2**.

- Prediction of pCR in HER2+ subtype:

Figure 2A shows the results for patients with HER2+ tumors, where Breast-NEOprAIdict achieved AUCs from 66.4% to 79.5% (worse to best partition). The CM achieved AUCs from 41.6% to 68.3%. Using the positivity threshold defined in the Methods section, Breast-NEOprAIdict achieved an OR of 4.44 (95% CI 2.11 - 9.38, *P* <0.0001), whereas the CM achieved an OR of 2.91 (95% CI 0.12 - 72.75, *P* = 1) (**Table 2**). Note that Breast-NEOprAIdict achieved an AUC from 0.60 to 0.89 on ER-/HER2+ subgroup, and an AUC from 0.59 to 0.73 on ER+/HER2+ subgroup.

- Prediction of pCR in ER+/HER2- subtype:

As shown in **Figure 2B** for the luminal breast cancer subtype (ER+/HER2-), Breast-NEOprAIdict achieved AUCs from 84.0% to 89.4%. In contrast, the CM achieved AUCs from 39.1% to 52.1% . Breast-NEOprAIdict achieved an OR of 72.53 (95% CI 4.36 - 1205.43, *P* <0.0001), whereas the CM achieved an OR of 0.319 (95% CI 0.0127 - 8.05, *P* = 1) (**Table 2**).

- Prediction of pCR in the Triple Negative Breast cancer subtype:

For patients with TN tumors (**Figure 2C**), Breast-NEOprAIdict achieved AUCs from 49.5% to 68.3%. The CM achieved similar AUCs from 51.5% to 68.6%. However, considering OR instead of AUC, Breast-NEOprAIdict achieved 2.22 (95% CI 1.23 - 4.00, *P* = 0.008) compared to 0.596 (95% CI 0.252 - 1.41, *P* = 0.259) for the CM (**Table 2**). This suggests that although Breast-NEOprAIdict is not better than the CM at any threshold, a careful threshold selection procedure shows its advantage over the CM. This is illustrated by the consistent shape of the Breast-NEOprAIdict ROC curve while, for the CM, the ROC curve with 68.6% could be an outlier.

Overall, the Breast-NEOprAIdict showed better performance in terms of both AUC and OR than the Clinical Model results, demonstrating the relevance of our approach to predict pCR versus RD in HER2+, ER+/HER2- and TN molecular subtypes.

External validation of Breast-NEOprAIdict on the CGFL Breast Cancer Neoadjuvant database

To validate the prediction performance of Breast-NEOprAIdict, we conducted an external validation on the CGFL breast cancer neoadjuvant database for a total of 490 patients (**Supplementary Figure S1B**). We used these thresholds for binarization (HER2+ = 0.38, ER+/HER2- = 0.41, TN = 0.50), following the binarization procedure described in the Methods. Breast-NEOprAIdict (**Figure 3A**) demonstrated high performance in predicting pCR with an AUC of 65.2% (*P* = 0.001) and an OR of 2.70 (95% CI 1.08-6.76, *P* = 0.0358, **Table 2**) for HER2+ tumors, an AUC of 81.4% (*P* = 0.003). Note that taken separately, ER-/HER2+ obtained an AUC of 0.58 and ER+/HER2+ an AUC of 0.67, suggesting that most of the discriminative effect is independent of the ER status for HER2+ tumors. Breast-NEOprAIdict obtained an OR of 20.56 (95% CI 1.14-371.74, *P* = 0.00413, **Table 2**) for

ER+/HER2- tumors, and an AUC of 67.7% (*P* = 0.001) and an OR of 3.02 (95% CI 1.18-7.74, *P* = 0.0206) for TN tumors (**Figure 3A**, **Table 2**). Although the number of pCR in the ER+/HER2- subgroup is only 6, a p-value of 0.004 suggests that the effect is likely to be strong. As an element of comparison, OncotypeDX and MammaPrint respectively achieve an OR of 4.48 and 2.25[45]. NPV for TN, HER2+ and ER+/HER2- are respectively 0.829, 0.891 and 1.00, other metrics such as sensitivity, specificity, PPV are provided in **Supplementary Table S3**. As shown in **Figure 3B** and **Table 2**, CM predicted pCR with an AUC of 51.8% (*P* = 0.353) and an OR of 3.63 (95% CI 0.07-185.45, *P* = 1) for HER2+ tumors, an AUC of 62.3% (*P* = 0.255) and an OR of 0.436 (95% CI 0.0237-8.01, *P* = 0.595) for ER+/HER2- tumors, and an AUC of 60.9% (*P* = 0.032) and an OR of 2.03 (95% CI 0.57-7.15, *P* = 0.308) for TN subtypes. These results demonstrate that Breast-NEOprAIdict, with a very strong NPV, is very effective in identifying the most chemoresistant tumors in patients for all molecular subtypes, which will result in the presence of RD after chemotherapy. As patient from the external validation has varying chemotherapy regimen (i.e. Taxanes, Anthracyclines or both Taxanes + Anthracyclines), we evaluated the predictive performance for each regiment when it was possible (**Supplementary Figure S7**). We found no major effect of the chemotherapy regimen on the pCR predictability.

Additionally, we highlight the advantages of combining the EfficientNet B7-based model with the ViT-S/16-based model. We present the results of the individual models in the external validation in **Supplementary Table S4**, showing that averaging the predictions from both models enhances performance across all molecular subtypes.

Comparing Breast-NEOprAIdict external validation predictions in patients with no-response, partial response, and complete response

The definition of pathological complete response (pCR) is limited to patients classified as pT0N0. Consequently, this categorization lumps together patients who show any degree of tumor regression and those who exhibit no improvement at all. We further stratified the predicted outcomes to separate non responders, partial and complete responder patients, as defined in the Materials and Methods section. As depicted in **Supplementary Figure S5**, there is an association between the prediction scores and the degrees of response to chemotherapy for patients with HER2+, ER+/HER2- and TN tumors. Difference in prediction score between partial responders and complete responders is significant for all molecular subtypes. However, this is not the case between non responders and partial responders, potentially because the model has not been trained to discriminate against them.

<u>Aggregation of the Breast-NEOprAIdict Model and the Clinical Model predictions show only minor improvement for TN tumors:</u>

We next tested whether combining Breast-NEOprAIdict and CM by averaging their slide-level prediction could lead to performance improvement, hereafter referred to as Breast-NEOprAIdict+CM. We compared the Breast-NEOprAIdict+CM model to the Breast-NEOprAIdict alone and found consistent improvement for the TN molecular subtype. When applied to the internal validation scheme, the Breast-NEOprAIdict+CM showed improvement in 3 out of 4 partitions over the Breast-NEOprAIdict alone. On the external validation, the improvement was a 1-point increase in performance (AUC of 68.7, P = 0.001) compared to the Breast-NEOprAIdict alone (AUC of 67.7, P = 0.001).

However, results are mixed for HER2+ and ER+/HER2- subtypes where internal validation tends to show an improvement, whereas external validation shows a 0.4 and 2.9 points reduction when compared with the Breast-NEOprAIdict alone for HER2+ and ER+/HER2- subtypes respectively. These findings suggest that while the Breast-NEOprAIdict Model is effective using only Whole Slide Image information for HER2+ and ER+/HER2- molecular subtypes, incorporating clinical information might benefit patients with TN tumors.

Breast-NEOprAIdict is consistent across different patient tissue samples:

To further evaluate the robustness of our method, we hypothesized that a reliable model should predict similar outputs across multiple biopsy slices from the same patient. Using the Breast-NEOprAIdict from the external validation scheme, we predicted the pCR status for 842 slides, originating from 421 patients in the external validation set (i.e. CGFL breast cancer neoadjuvant database), with each patient providing two slide images from two distinct biopsy sections.

Our method demonstrated a strong positive correlation between predictions for a same patient, as shown by a Pearson correlation coefficient of 0.933 (P = 5.78e-90/<0.001) for HER2+, 0.932 (P = 4.51e-57/<0.001) for ER+/HER2-, and 0.939 (P = 5.24e-44/<0.001) for TN molecular subtypes. Similarly, the binary concordance measured with the Kappa coefficient showed a high degree of agreement, with a value of 0.711 for HER2+, 0.728 for ER+/HER2- and 0.929 for TN molecular subtypes (**Figure 4**). These results show that our model is consistent in predicting pCR across multiple biopsies from the same patient.

Contrasting Morphological Features in Chemosensitive vs Chemoresistant Tumors: Insights from High and Low Scoring Slides by Breast-NEOprAIdict analyzed by pathologists:

Analysis of high score patches by experts revealed visually discernible features considered by the Breast-NEOprAIdict algorithm to be representative of sensitivity or resistance to chemotherapy, providing some insights into the predictive capabilities of our model.

For HER2+ tumors (**Figure 5A**), the patches identified by Breast-NEOprAIdict as the most chemosensitive exhibit a high cell density, large nucleolated nuclei, inflammatory stroma, lymphocytic infiltration, and numerous mitoses. In contrast, the patches identified as the most chemoresistant are sparsely cellular, forming clusters with cord-like patterns, surrounded by fibrous stroma and a low lymphocytic density. Among luminal tumors (**Figure 5B**), patches with the highest scores in the slides considered as the most chemosensitive by the algorithm, regardless of the ground truth, displayed a consistent nucleocytoplasmic ratio with prominent clear nucleoli and heterogeneous chromatin[21]. In contrast, patches from tumors deemed the most chemoresistant by the algorithm, i.e. the slides with the lowest scores, exhibited small nuclei with homogeneous chromatin, low tumor density, and collagen-rich stroma. TILs were present in both categories, suggesting they do not serve as a discriminant feature for the neural network.

In TN tumors (**Figure 5C**), patches with the highest scores for the tumors considered as the most chemosensitive showed high cell density, lymphocyte-rich stroma and significant pleomorphism. Conversely, patches from tumors considered as the most chemoresistant displayed collagen-rich stroma outlining tumor islands with a cordonal arrangement.

Discussion

In this study, we introduced Breast-NEOprAIdict, a new DL-based tool to predict pCR after NAC using WSI of initial diagnostic biopsies. We drew on two cohorts of eBC patients undergoing NAC treatment, collectively consisting of 1140 selected patients from 12 distinct cancer centers throughout France. The PRIMUNEO cohort, spanning the years 2012 to 2015, had a higher pCR rate (24%) compared to the CGFL cohort (17.9%). This disparity is attributed to the more contemporary nature of the PRIMUNEO dataset and the overall improved performance of chemotherapy regimens in recent years. The CGFL dataset also featured a broader array of treatment protocols compared to those in the PRIMUNEO dataset, which is a logical outcome given its basis in 'real life data', outside the setting of a clinical trial, and its extended duration of patient follow-up.

Our study introduces several pioneering elements: 1) we propose a pan-molecular subtype analysis and pCR prediction among the biological variety of eBC on a vast multicentre cohort; 2) we affirm the predictive capacity of our deep learning approach using an independent cohort; 3) to the best of our knowledge, this is the first attempt to propose an analysis of the robustness of AI algorithms in Digital Pathology by systematically evaluating the concordance of the Deep Learning predictions on

different biopsy sections from the same patient; and finally 4) we provide new information about the pathological features associated with chemosensitivity of luminal and triple-negative tumors discovered by our algorithm.

Breast cancer is a multifaceted and diverse condition necessitating tailored treatment strategies based on the characteristics of both the tumor and the patient. NAC not only increases the rate of conservative surgery, but also facilitates *in vivo* assessment of the chemotherapy sensitivity of the tumor. It also makes it possible to propose an adjuvant systemic treatment strategy for tumors that have not responded to a standard NAC protocol. Nevertheless, NAC can also be associated with significant side effects, making it essential to identify patients who would benefit the most from it. For example, in ER+/HER2- eBC, clinicians could reserve chemotherapy for the small fraction of chemosensitive, clinical high-risk luminal tumors, and orient other patients towards intensified adjuvant endocrine therapy (with CDK4/6 inhibitors for example, or any investigating new drugs)[22]. In TN breast cancer, clinicians could avoid systematic neoadjuvant therapeutic escalation, such as carboplatin[23] and immunotherapy with anti PD-1[24] (which are the current standard of care for stage II-III eTNBC), and unnecessary cytotoxicity in patients who are likely to achieve pCR with standard sequential anthracycline-taxane chemotherapy[25,26] (such as the patients analyzed in our study). It is noteworthy to emphasize that our results are derived from patients treated without carboplatin or immunotherapy. This is particularly relevant for de-escalation strategies aimed at identifying tumors that are sensitive to standard protocols, potentially sparing certain patients from unnecessary intensified treatments.

Since the initial development of NAC in eBC, numerous clinical biomarkers (tumor size or initial clinical stage[27]) alone, or combined with biological factors (HR, HER2 or Ki67 expression levels, tumor grade, TILs[28–31]) have been reported to be associated with the probability of pCR achievement. More recently, some gene-expression signatures[32], but also radiomic features in MRI[33,34] or PET imaging[35], evaluated either at baseline or early after initiation of NAC, have also been developed to predict pCR. However, these methods have not yet gained widespread acceptance in everyday clinical practice for diagnosing patients and suggesting treatments based on expected tumor responsiveness to chemotherapy.

In recent years, deep learning algorithms have shown promising results for the prediction of various information using Whole Slide Imaging (WSI), including classical histological features[36], genomic mutations[37] and survival prediction[38]. These tools have the ability to provide fast, low-cost, and accurate predictions for a wide range of applications and these (H&E) WSI are beginning to be used routinely in the diagnostic workflow. A deep learning-based solution predicting pCR directly from the initial H&E biopsies has been proposed by other teams as a potentially valid predictive strategy[9,39–41].

Most of these pioneering works[9,39,40] focused on pCR prediction in TNBC, using Deep Learning and combining H&E slides with IHC slide analyses, or used more standard machine learning approaches[41]. Our study aimed to predict pCR in breast cancer patients who received standard NAC by developing and validating a deep learning system using initial diagnostic tumor biopsies, denoted as Breast-NEOprAIdict. Our key discovery is that the algorithm notably outperformed traditional models based on clinical and pathological markers, particularly in predicting pCR in patients with HER2-amplified, ER-positive/HER2-negative and triple-negative (TN) breast cancer in large independent cohorts. Indeed, our method outperformed the Clinical Model (CM), which was designed to combine clinical information such as the age, molecular subtype, tumor grade, and clinical tumor burden (estimated by cAJCC staging) for predicting pCR status in eBC patients. Moreover, combining these gold standard clinico-pathological markers with Breast-NEOprAIdict did not improve the performance of our solution in patients with HER2+ and ER+/HER2- tumors showing that Breast-NEOprAIdict could be used alone to predict response to chemotherapy, however this usual clinicopathological factors could benefit the response to chemotherapy prediction in patients with TN tumors. The very high negative predictive value of our solution on HER2+ (NPV=87.2%),, luminal (NPV=100%), and triple negative (NPV=80.0%) tumors could enable clinicians to select only a subset of patients who would benefit from NAC intensification (carboplatin and immunotherapy in TNBC, endocrine-based treatment in ER+/HER2- subtypes, or new antibody-drug conjugates (ADCs) targeting HER2 in HER2-amplified BC for example).

These findings are congruent with existing literature underscoring the potential of histological images in forecasting treatment response[9]. However, unlike previous studies, which were confined to more traditional machine learning approaches[41] or specific subtypes like TNBC[40], our research casts a wider net by embracing all molecular categories of breast cancer, thereby contributing a new, comprehensive perspective to the field.

One of the important points of our report, and a unique contribution based on our understanding of the existing literature, lies in our unprecedented ability within the field of digital pathology to assess the biological concordance of our method. We systematically tested the consistency of our predictions by applying them to distinct samples for each patient. Our approach unveiled consistent predictive capability in different samples from the same patient, a step forward in confirming the model's reliability against data variability. Our investigation extended to analyzing the predictions generated by our deep learning tool, aided by the expertise of seasoned pathologists. Interestingly, the areas that Breast-NEOprAIdict highlighted as key to distinguishing chemosensitive from chemoresistant tumors match findings from pathology studies. Breast-NEOprAIdict naturally identifies known features linked to chemosensitivity, like lymphocyte-rich stroma in TN tumors (the

most chemosensitive), and collagen-rich stroma, which is linked to the most chemoresistant tumors[42,43]. In addition, our Breast-NEOprAIdict has brought to light new morphometric parameters that are important for predicting the chemosensitivity of luminal tumors, such as those inherent in nuclear and nucleolar morphology. Consistent with existing literature in ER+/HER2- eBC[44], TILs do not appear to serve as a discriminant feature for the neural network.

Despite these advances, our study acknowledges several limitations, mostly due to the lack of multiple centers for the external validation and because the ER+/HER2- subtype had only 6 patients with pCR in the external validation set. However, one notable quality of our data is the large period it spans (more than 20 years). The overall good biological concordance of our model could also be improved. Deeper investigation of the cases where patient-based predictions did not align could be highly significant. Additionally, our dataset's bias towards certain molecular subtypes and reliance on a single type of scanner for all WSIs might constrain the model's broader applicability. These limitations underscore the necessity for caution in interpreting our findings and their implications in real-world clinical settings.

Given these insights and constraints, future research should pursue several avenues. Firstly, it is imperative to validate these findings through prospective multicenter studies, encompassing a wider representation of HER2+ tumors and using varied scanning technologies. Finally, considering survival data as a stratification parameter might present a more nuanced understanding of treatment responses over time, potentially reshaping how we perceive and use pCR as a solitary metric for therapeutic success.

In conclusion, our study marks a significant step towards personalized breast cancer treatment, demonstrating the potential of deep learning models to predict pCR with notable accuracy in certain breast cancer subtypes. By enabling the distinction between chemosensitive and chemoresistant tumors, our findings facilitate more informed, individualized therapeutic decision-making. However, the road ahead requires comprehensive, nuanced studies to refine these predictive tools, ensuring they are robust and versatile enough for diverse clinical scenarios. As we bridge this critical gap, the overarching vision is a future where every breast cancer patient receives optimized, effective therapy, significantly enhancing survival rates and quality of life.

References:

1. Cardoso, F. *et al.* Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 1194–1220 (2019).
2. Masood, S. Neoadjuvant chemotherapy in breast cancers. *Womens Health* **12**, 480–491 (2016).
3. Asselain, B. *et al.* Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. *Lancet Oncol.* **19**, 27–39 (2018).
4. Cortazar, P. *et al.* Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *The Lancet* **384**, 164–172 (2014).
5. Spring, L. M. *et al.* Pathologic Complete Response after Neoadjuvant Chemotherapy and Impact on Breast Cancer Recurrence and Survival: A Comprehensive Meta-analysis. *Clin. Cancer Res.* **26**, 2838–2848 (2020).
6. Masuda, N. *et al.* Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. *N. Engl. J. Med.* **376**, 2147–2159 (2017).
7. Von Minckwitz, G. *et al.* Trastuzumab Emtansine for Residual Invasive HER2-Positive Breast Cancer. *N. Engl. J. Med.* **380**, 617–628 (2019).
8. Wolff, A. C. *et al.* Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Arch. Pathol. Lab. Med.* **142**, 1364–1382 (2018).
9. Li, F. *et al.* Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer. *J. Transl. Med.* **19**, 348 (2021).
10. Amat, S. *et al.* Scarff-Bloom-Richardson (SBR) grading: a pleiotropic marker of chemosensitivity in invasive ductal breast carcinomas treated by neoadjuvant chemotherapy. *Int. J. Oncol.* (2002) doi:10.3892/ijo.20.4.791.
11. Giuliano, A. E., Edge, S. B. & Hortobagyi, G. N. Eighth Edition of the AJCC Cancer Staging Manual: Breast Cancer. *Ann. Surg. Oncol.* **25**, 1783–1785 (2018).
12. Howard, F. M. *et al.* The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
13. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2019) doi:10.48550/ARXIV.1905.11946.
14. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at http://arxiv.org/abs/2010.11929 (2021).
15. Caron, M. *et al.* Emerging Properties in Self-Supervised Vision Transformers. Preprint at https://doi.org/10.48550/ARXIV.2104.14294 (2021).
16. Foret, P., Kleiner, A., Mobahi, H. & Neyshabur, B. Sharpness-Aware Minimization for Efficiently Improving Generalization. Preprint at http://arxiv.org/abs/2010.01412 (2021).
17. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
18. Lawson, R. Small Sample Confidence Intervals for the Odds Ratio. *Commun. Stat. - Simul. Comput.* **33**, 1095–1113 (2004).
19. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426219/.
20. Nemes, S., Jonasson, J. M., Genell, A. & Steineck, G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med. Res. Methodol.* **9**, 56 (2009).
21. Delahunt, B. *et al.* Gleason and Fuhrman no longer make the grade. *Histopathology* **68**, 475–481 (2016).
22. Chien, A. J., Kyalwazi, B. & Esserman, L. J. Optimizing hormone therapy for breast cancer: Translating gains to the early-stage setting. *Cell Rep. Med.* **3**, 100664 (2022).

23. Geyer, C. E. *et al.* Long-term efficacy and safety of addition of carboplatin with or without veliparib to standard neoadjuvant chemotherapy in triple-negative breast cancer: 4-year follow-up data from BrighTNess, a randomized phase III trial. *Ann. Oncol.* **33**, 384–394 (2022).
24. Schmid, P. *et al.* Event-free Survival with Pembrolizumab in Early Triple-Negative Breast Cancer. *N. Engl. J. Med.* **386**, 556–567 (2022).
25. Shah, A. N. *et al.* Phase II study of pembrolizumab and capecitabine for triple negative and hormone receptor-positive, HER2−negative endocrine-refractory metastatic breast cancer. *J. Immunother. Cancer* **8**, e000173 (2020).
26. Robson, M. *et al.* Olaparib for Metastatic Breast Cancer in Patients with a Germline *BRCA* Mutation. *N. Engl. J. Med.* **377**, 523–533 (2017).
27. Goorts, B. *et al.* Clinical tumor stage is the most important predictor of pathological complete response rate after neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Res. Treat.* **163**, 83–91 (2017).
28. Lips, E. H. *et al.* Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response. *Breast Cancer Res. Treat.* **140**, 63–71 (2013).
29. Pu, S. *et al.* Nomogram-derived prediction of pathologic complete response (pCR) in breast cancer patients treated with neoadjuvant chemotherapy (NCT). *BMC Cancer* **20**, 1120 (2020).
30. Kim, S.-Y. *et al.* Factors Affecting Pathologic Complete Response Following Neoadjuvant Chemotherapy in Breast Cancer: Development and Validation of a Predictive Nomogram. *Radiology* **299**, 290–300 (2021).
31. Li, S. *et al.* Predictive and prognostic values of tumor infiltrating lymphocytes in breast cancers treated with neoadjuvant chemotherapy: A meta-analysis. *The Breast* **66**, 97–109 (2022).
32. Griguolo, G. *et al.* Gene-expression signatures to inform neoadjuvant treatment decision in HR+/HER2− breast cancer: Available evidence and clinical implications. *Cancer Treat. Rev.* **102**, 102323 (2022).
33. Liu, Z. *et al.* Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study. *Clin. Cancer Res.* **25**, 3538–3547 (2019).
34. Eun, N. L. *et al.* Texture Analysis with 3.0-T MRI for Association of Response to Neoadjuvant Chemotherapy in Breast Cancer. *Radiology* **294**, 31–41 (2020).
35. Coudert, B. *et al.* Use of [18F]-FDG PET to predict response to neoadjuvant trastuzumab and docetaxel in patients with HER2-positive breast cancer, and addition of bevacizumab to neoadjuvant trastuzumab and docetaxel in [18F]-FDG PET-predicted non-responders (AVATAXHER): an open-label, randomised phase 2 trial. *Lancet Oncol.* **15**, 1493–1502 (2014).
36. Courtiol, P., Tramel, E. W., Sanselme, M. & Wainrib, G. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. *ArXiv180202212 Cs Stat* (2020).
37. Morel, L.-O., Derangère, V., Arnould, L., Ladoire, S. & Vinçon, N. Preliminary evaluation of deep learning for first-line diagnostic prediction of tumor mutational status. *Sci. Rep.* **13**, 6927 (2023).
38. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
39. Duanmu, H. *et al.* A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images. *Bioinformatics* **38**, 4605–4612 (2022).
40. Naylor, P. *et al.* Prediction of Treatment Response in Triple Negative Breast Cancer From Whole Slide Images. *Front. Signal Process.* **2**, 851809 (2022).
41. Ogier du Terrail, J. *et al.* Collaborative federated learning behind hospitals' firewalls for predicting histological complete response to neoadjuvant chemotherapy in triple-negative breast cancer. *J. Clin. Oncol.* **40**, 590–590 (2022).

42. Smith, J. *et al.* Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: The Impact on Pathological Complete. *J. Clin. Oncol.* **38**(15), 1579-1588 (2020).
43. Patel, A. & Gupta, R. The Emerging Role of Pathology AI in Breast Cancer Prognosis and Treatment. *Pathol. Today* **12**, 245–251 (2019).
44. Ali HR, *et al.* Association between CD8+ T-cell infiltration and breast cancer survival in 12,439 patients. Ann Oncol. 2014 Aug;**25**(8):1536-43.
45. Freeman, J. Q. et al. Evaluation of multigene assays as predictors for response to neoadjuvant chemotherapy in early-stage breast cancer patients. npj Breast Cancer 9, 1–4 (2023).

Conflict of interest

Louis-Oscar Morel and Nathan Vinçon own shares in the Ummon Healthtech company.

Code Availability

Source code is publicly available at Breast-NEOprAIdict repository in https://gitlab.com/nfvalderrama/breast-neopraidict.

Author contribution

LOM, SL and NV conceived the study. LOM and NFV wrote the article. LOM and NFV performed the analysis and interpretation. LA and HPM analyzed the patches coming from the WSI biopsies. LA and VD did and coordinated all the reports coming from the CGFL. DTM digitized, collected and assembled the biopsies coming from the CGFL. ID, DM, CK, SI, AH, NR, and AB contributed to the PRIMUNEO study and dataset development. NV and SL reviewed the final manuscript.

Figure Legends

Figure 1. Pipeline Overview.

Breast-NEOprAIdict predicts patient response to neoadjuvant therapy using Whole Slide Images (WSIs). Our approach: A) extracts the foreground from WSIs, tiling tissue into patches, and using an EfficientNetB7 neural network pre-trained on ImageNet; B) to compute 2560-dimensional features. These features undergo analysis through a Multilayer Perceptron (MLP), incorporating molecular subtype information. C) The final prediction for pathological complete response (pCR) or residual

disease (RD) and molecular subtype is obtained through a SoftMax layer. The combined score is normalized, ensuring appropriate weighting based on clinical information. Slide-level predictions are determined by the 99th percentile of patch-level prediction values.

Figure 2. Comparison of AUC between Breast-NEOprAIdict and Clinical model (CM) in internal validation.

Receiver operating characteristic (ROC) curves illustrate the performance of Breast-NEOprAIdict (left) and CM (right) across various partitions for A) HER2+, B) ER+/HER2- and C) TN breast cancer molecular subtypes. Legends in the figures display the Area Under the Curve (AUC) and p-value (P).

Figure 3. Breast-NEOprAIdict vs Clinical model (CM) AUC performance comparison in the external validation.

The ROC curves illustrate the performance of (A) Breast-NEOprAIdict and (B) CM across the different molecular subtypes of tumors. Legends in the figures provide details on the Area Under the Curve (AUC) and associated p-values (P).

Figure 4. Agreement in pCR score prediction between two different slides from the same patient.

pCR prediction on a patient's slide versus prediction on a second slide from the same patient. Pearson's correlation coefficient (r) and Kappa concordance score (k) are shown. The figure indicates the threshold used for dichotomizing the prediction (Binary thr), and the color labels indicate whether the binary prediction within the patient slides agreed or not (green for yes, and purple for no).

Figure 5. Slide visual analysis.

Comparison of the patches with the highest scores in the slides considered as chemosensitive (left, highest overall scores) or chemoresistant (right, lowest overall scores). Part A) displays patches coming from HER2+, Part B) ER+/HER2- and Part C) from TN tumors.

Supplementary Figure S1. Workflow for patient selection in the PRIMUNEO and CGFL Breast Cancer Neoadjuvant databases. The diagram illustrates the patient selection process in the PRIMUNEO and CGFL Breast Cancer Neoadjuvant databases. A) The chart specifies the included patients in PRIMUNEO to compose the internal validation setup. B) The training set for the external validation setup is then derived by excluding patients from the CGFL center in Dijon in the PRIMUNEO cohort.

The test set is composed by the CGFL Breast Cancer Neoadjuvant database. The chart considers the number of patients with pCR or RD categorized by the tumor's molecular subtype.

Supplementary Table S1. Description of the PRIMUNEO Database, including the number of patients exhibiting either pathological complete response (pCR) or residual disease (RD), categorized by the tumor's molecular subtype within each center.

Supplementary Table S2. CGFL breast cancer neoadjuvant database description according to tumor molecular subtype.

Supplementary Figure S2. The chart shows the specifications of the laboratory, total number of patients and pCR rate (%) in each partition for the training and test subsets in the internal validation study. The distribution of the patients is also shown by molecular subtype. Best viewed in color.

Supplementary Figure S3. Three-Fold Cross-Validation training strategy.

Supplementary Figure S4. Prediction ensemble for computation of OR metrics.

Supplementary Table S3. Breast-NEOprAIdict performance is measured with sensitivity (recall), PPV positive predictive value (precision), specificity, and NPV negative predictive value on external validation for HER2+, ER+/HER2- and TN breast cancer molecular subtypes.

Supplementary Figure S5. Comparison of Breast-NEOprAIdict score prediction distribution within non-responder, partial responder, and complete responder patients with A) HER2+, B) ER+/HER2- and C) TN tumors. The one-sided Mann-Whitney-Wilcoxon test was used to compare the response categories. ns: 0.05 < p-value <= 1; *: 0.01 < p-value <= 0.05; **: 0.001 < p-value <= 0.01, ***: 0.0001 < p <= 0.001.

Supplementary Figure S6. Evaluation of Breast-NEOprAIdict pCR prediction performance in all patients compared to those who did not respond or exhibited complete response to Neoadjuvant Chemotherapy (excluding patients with partial response defined in the refined labels). We show the results for patients with A) HER2+, B) ER+/HER2- and C) TN tumors in external validation. Legends in the figures provide details on the Area Under the Curve (AUC) and associated p-values (P).

Supplementary Figure S7. AUC in patients with A) HER2+, B) ER+/HER2- and C) TN tumors in the external validation study, stratifying patients by the received treatment (Taxanes, Anthracyclines, or both).

Supplementary Figure S8. AUC in patients with A) HER2+, B) ER+/HER2- and C) TN tumors in the internal validation study. The patients' pCR predictions are obtained by averaging Breast-NEOprAIdict and CM predictions.

Supplementary Figure S9. AUC in patients with HER2+, ER+/HER2-, and TN tumors in the external validation study. The patients' pCR predictions are obtained by averaging Breast-NEOprAIdict and CM predictions.

Supplementary Table S4. Ablation study on the ensemble of the EfficientNet B7-based model and the ViT-S/16-based model in the external validation. AUC in patients with HER2+, ER+/HER2-, and TN tumors is shown.

Table 1. PRIMUNEO and CGFL breast cancer neoadjuvant dataset characteristics.

Variables	PRIMUNEO (N=500)	CGFL breast cancer neoadjuvant database (N=640)	P value
Age, median, range	50 [22 - 78]	51 [23 - 88]	0.4399
Menopausal status			0.2013
Premenopausal	269 (57.0%)	312 (53.1%)	
Postmenopausal	203 (43.0%)	276 (46.9%)	
Missing	28	52	
cT stage			< 0.001
T0	2 (0.4%)	2 (0.4%)	
T1	33 (6.9%)	49 (8.5%)	
T2	269 (56.2%)	374 (64.9%)	
T3	117 (24.4%)	58 (10.1%)	
T4	58 (12.1%)	93 (16.1%)	
Missing	21	64	
cN stage			< 0.001
N0	195 (43.3%)	214 (37.3%)	

N1	226 (50.2%)	239 (41.6%)	
N2	16 (3.6%)	51 (8.9%)	
N3	13 (2.9%)	70 (12.2%)	
Missing	50	66	
SBR			< 0.001
1	13 (2.7%)	56 (9.4%)	
2	191 (39.5%)	278 (46.6%)	
3	279 (57.8%)	262 (44.0%)	
Missing values	17	44	
Clincial tumor stage (cAJCC)			0.7967
I	17 (3.7%)	19 (3.3%)	
II	302 (65.1%)	365 (63.6%)	
III	145 (31.3%)	190 (33.1%)	
Missing	36	66	
Tumor Molecular Subtype			< 0.001
HER2+	131 (26.9%)	246 (42.0%)	
ER+/HER2-	206 (42.4%)	193 (33.0%)	
TN	149 (30.7%)	146 (25.0%)	
Missing	14	55	
Type of NAC			< 0.001
Others	1 (0.2%)	21 (3.6%)	
Taxanes	77 (17.8%)	113 (19.5%)	
Anthracyclines	2 (0.5%)	111 (19.1%)	
Anthracyclines and taxanes	405 (93.1%)	335 (57.7%)	
Missing	15	60	
Breast surgery			0.0526
Radical	232 (47.3%)	198 (41.0%)	
Conservative	259 (52.7%)	284 (59.0%)	
Missing	9	158	
ypT stage			0.0327
T0	114 (25.2%)	120 (22.6%)	
T1	196 (43.3%)	270 (50.9%)	

T2	99 (21.9%)	111 (20.9%)	
T3	36 (7.9%)	26 (5.0%)	
T4	8 (1.8%)	3 (0.6%)	
Missing	47	110	
ypN stage			0.0289
N0	276 (58.7%)	276 (52.8%)	
N1	132 (28.1%)	141 (26.9%)	
N2	47 (10.0%)	81 (15.5%)	
N3	15 (3.2%)	25 (4.8%)	
Missing	30	117	
pathologic tumor stage (pAJCC)			0.2459
0	110 (23.8%)	132 (23.4%)	
I	110 (23.8%)	153 (27.1%)	
II	159 (34.3%)	164 (29.0%)	
III	84 (18.1%)	116 (20.5%)	
Missing values	37	75	
pCR status			0.0180
No	357 (76%)	430 (82.1%)	
Yes	113 (24%)	94 (17.9%)	
Missing	30	116	

HER2, human epidermal growth factor; HER2+, HER2-positive; HER2-, HER2-negative; ER, estrogen receptor; ER+, estrogen receptor positive; TN, Triple Negative; SBR, Scarff Bloom Richardson grade; AJCC, American Joint Committee on Cancer; c, clinical; p, pathological; NAC, neoadjuvant chemotherapy; RT, radiotherapy; pCR, pathological Complete Response.

Figure 1.

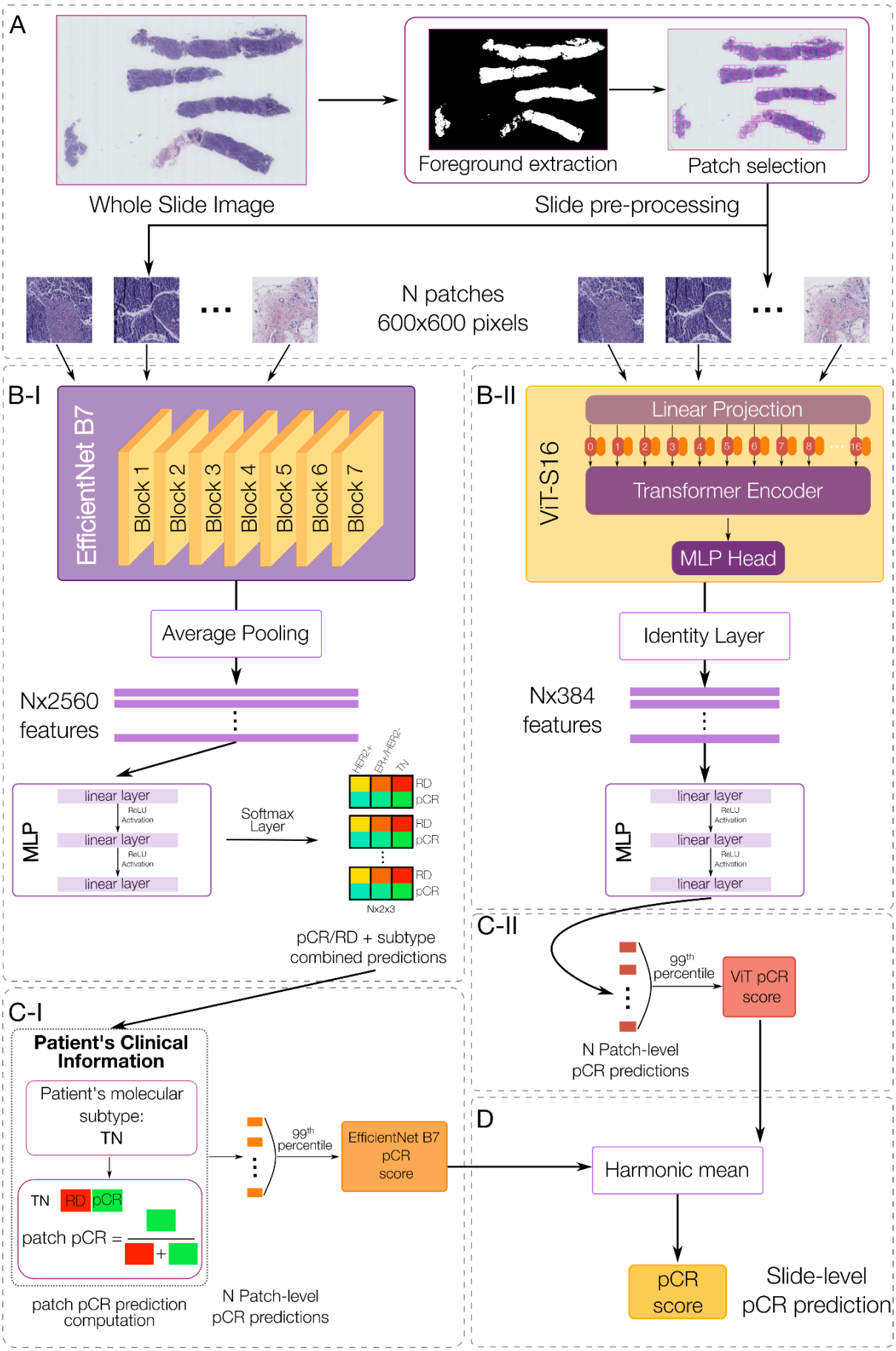

A
Whole Slide Image
Foreground extraction
Patch selection
Slide pre-processing
N patches
600x600 pixels
B-I
EfficientNet B7
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Average Pooling
Nx2560
features
MLP
linear layer
ReLU Activation
linear layer
ReLU Activation
linear layer
Softmax Layer
HER2+
ER+/HER2-
TN
RD
pCR
Nx2x3
pCR/RD + subtype
combined predictions
B-II
ViT-S16
Linear Projection
Transformer Encoder
MLP Head
Identity Layer
Nx384
features
MLP
linear layer
ReLU Activation
linear layer
ReLU Activation
linear layer
C-II
99th percentile
ViT pCR score
N Patch-level
pCR predictions
C-I
Patient's Clinical Information
Patient's molecular subtype:
TN
TN
RD
pCR
patch pCR =
patch pCR prediction computation
99th percentile
EfficientNet B7 pCR score
N Patch-level
pCR predictions
D
Harmonic mean
pCR score
Slide-level
pCR prediction

Figure 2.

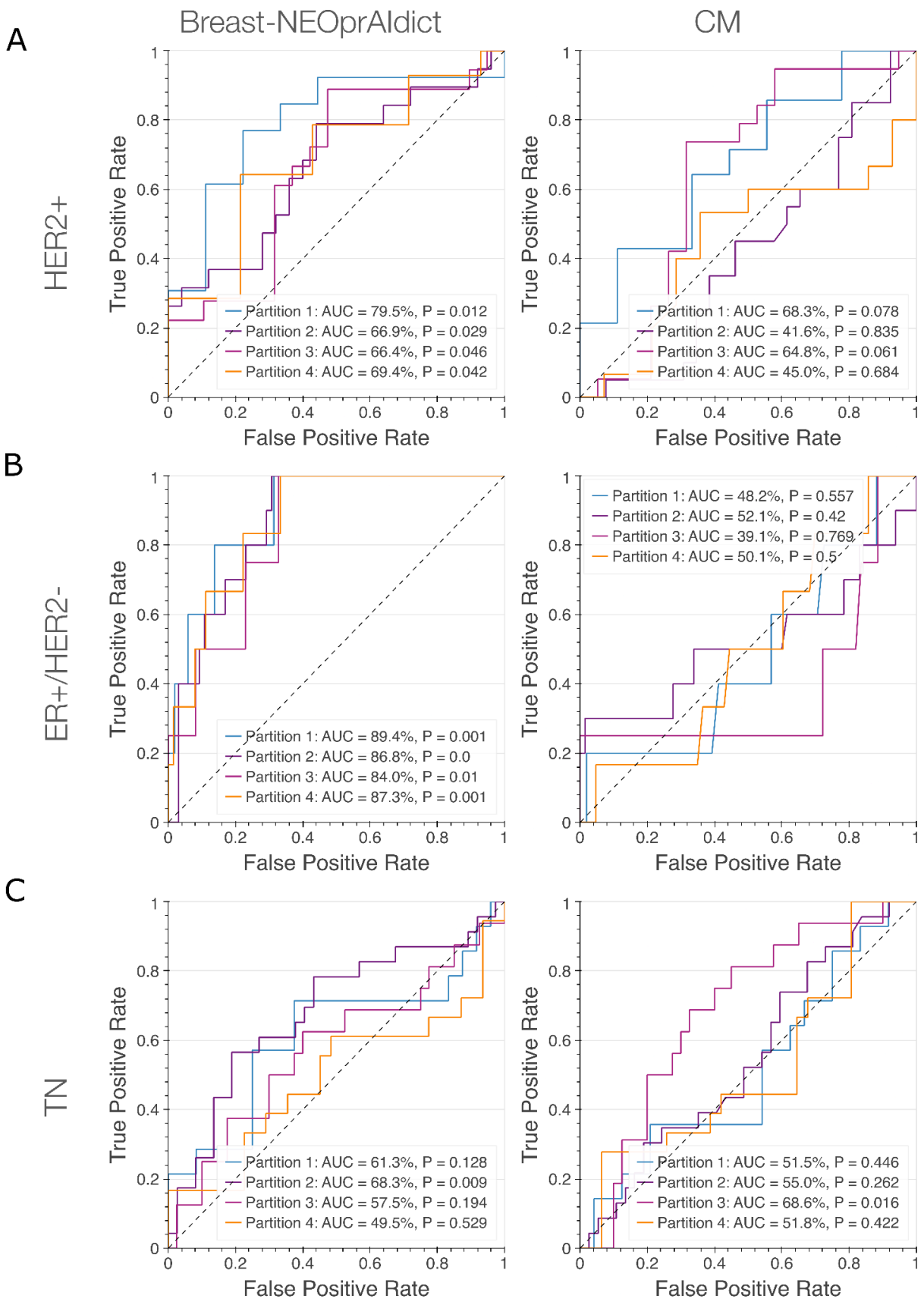

A
Breast-NEOprAIdict
CM
HER2+
True Positive Rate
False Positive Rate
Partition 1: AUC = 79.5%, P = 0.012
Partition 2: AUC = 66.9%, P = 0.029
Partition 3: AUC = 66.4%, P = 0.046
Partition 4: AUC = 69.4%, P = 0.042
Partition 1: AUC = 68.3%, P = 0.078
Partition 2: AUC = 41.6%, P = 0.835
Partition 3: AUC = 64.8%, P = 0.061
Partition 4: AUC = 45.0%, P = 0.684
B
ER+/HER2-
Partition 1: AUC = 89.4%, P = 0.001
Partition 2: AUC = 86.8%, P = 0.0
Partition 3: AUC = 84.0%, P = 0.01
Partition 4: AUC = 87.3%, P = 0.001
Partition 1: AUC = 48.2%, P = 0.557
Partition 2: AUC = 52.1%, P = 0.42
Partition 3: AUC = 39.1%, P = 0.769
Partition 4: AUC = 50.1%, P = 0.5
C
TN
Partition 1: AUC = 61.3%, P = 0.128
Partition 2: AUC = 68.3%, P = 0.009
Partition 3: AUC = 57.5%, P = 0.194
Partition 4: AUC = 49.5%, P = 0.529
Partition 1: AUC = 51.5%, P = 0.446
Partition 2: AUC = 55.0%, P = 0.262
Partition 3: AUC = 68.6%, P = 0.016
Partition 4: AUC = 51.8%, P = 0.422

Table 2. Comparison of ORs between Breast-NEOprAIdict and Clinical model (CM).

Validation Setup	Methods	Breast-NEOprAIdict		Clinical Model (CM)	
		OR (95 % CI)	***P***	**OR (95 % CI)**	***P***
Internal	HER2+	4.44 (2.11 - 9.38)	8.8e-05	2.91 (0.12-72.75)	1
	ER+/HER2-	72.53 (4.36-1205.43)	1.36e-09	0.319 (0.0127-8.05)	1
	TN	2.22 (1.23-4.00)	0.00831	0.596 (0.252-1.41)	0.259
External	HER2+	2.70 (1.08-6.76)	0.0358	3.63 (0.07-185.45)	1
	ER+/HER2-	20.56 (1.14-371.74)	0.00413	0.436 (0.0237-8.01)	0.595
	TN	3.02 (1.18-7.74)	0.0206	2.03 (0.57-7.15)	0.308

Figure 3.

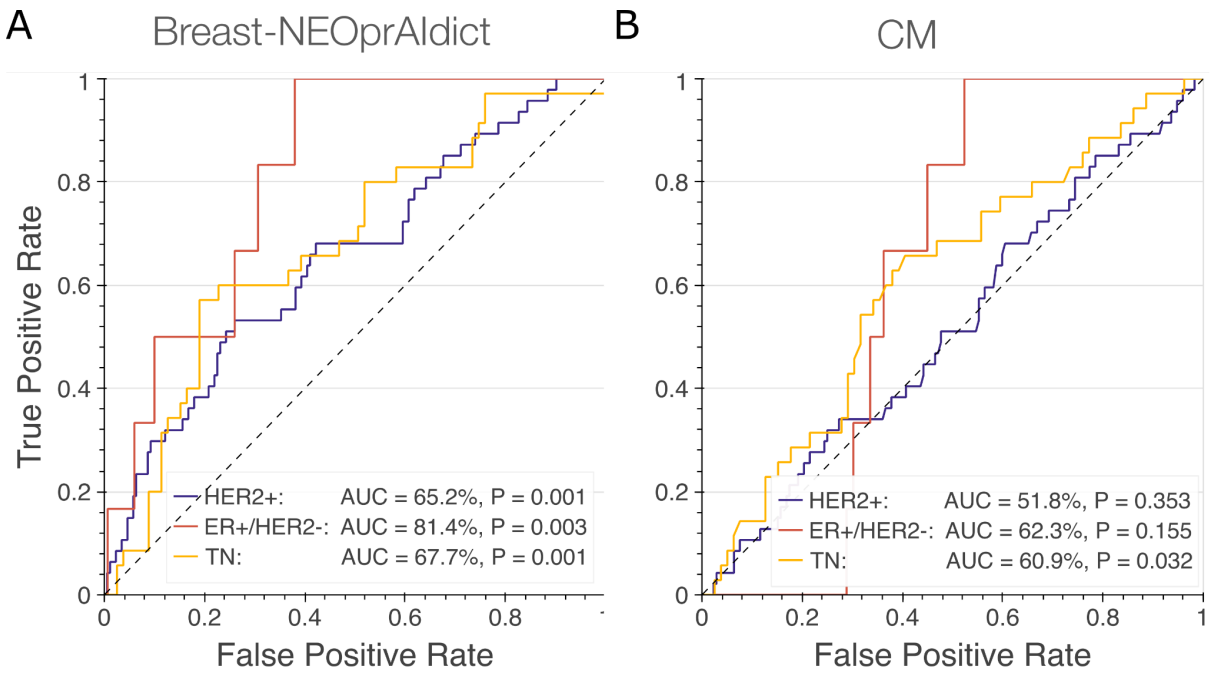

A
Breast-NEOprAIdict
True Positive Rate
False Positive Rate
HER2+: AUC = 65.2%, P = 0.001
ER+/HER2-: AUC = 81.4%, P = 0.003
TN: AUC = 67.7%, P = 0.001
B
CM
False Positive Rate
HER2+: AUC = 51.8%, P = 0.353
ER+/HER2-: AUC = 62.3%, P = 0.155
TN: AUC = 60.9%, P = 0.032

Figure 4.

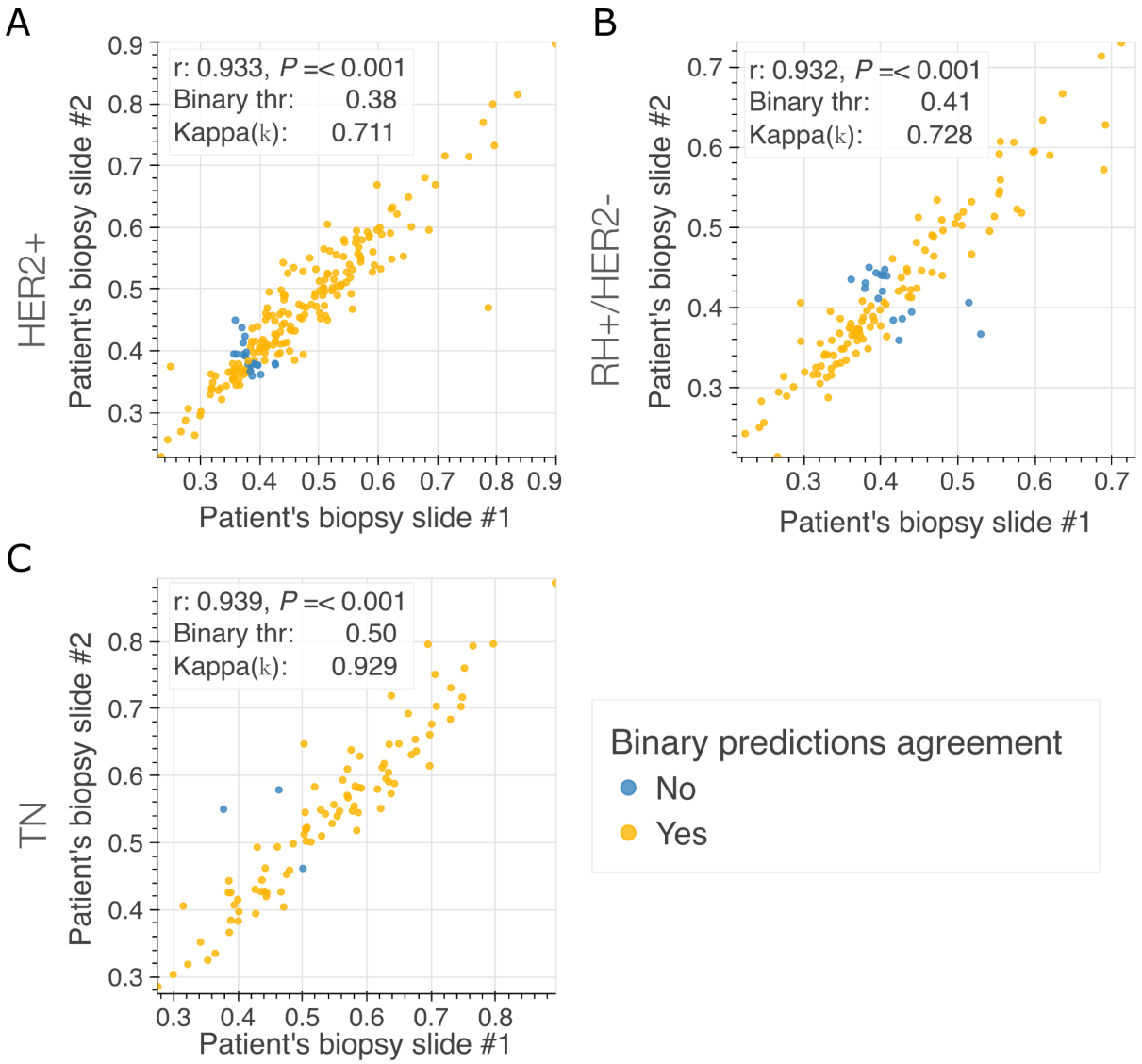
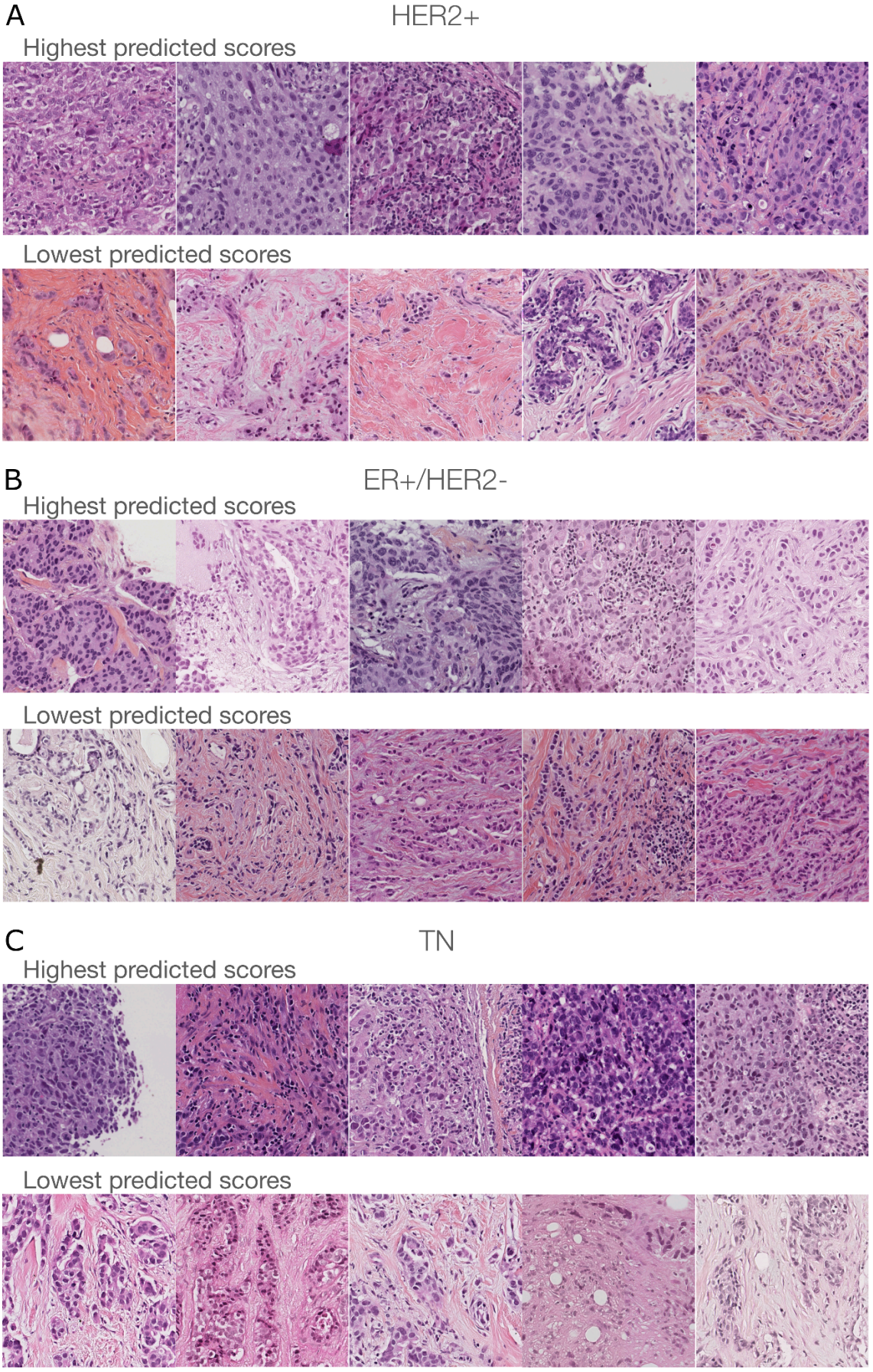
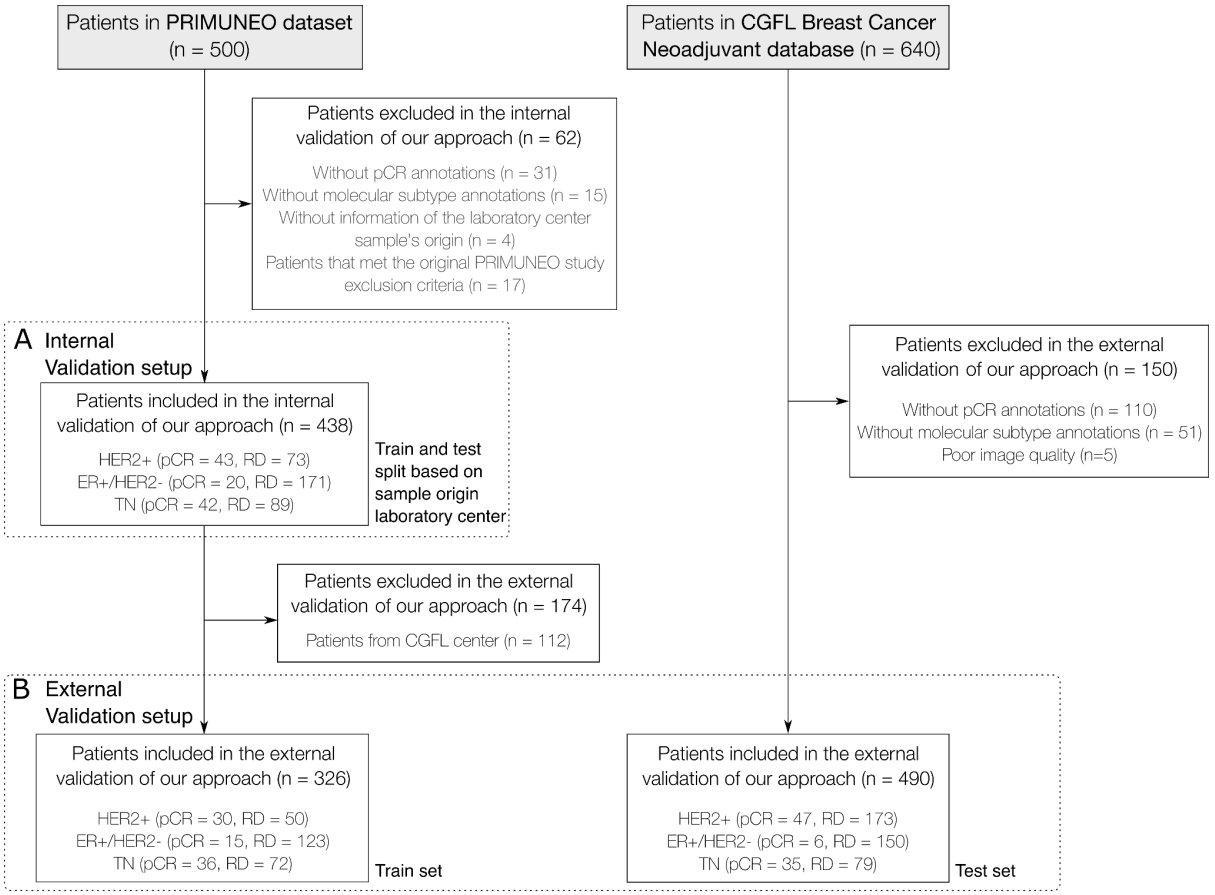

A
HER2+
Patient's biopsy slide #2
Patient's biopsy slide #1
r: 0.933, P =< 0.001
Binary thr: 0.38
Kappa(k): 0.711
B
RH+/HER2-
Patient's biopsy slide #2
Patient's biopsy slide #1
r: 0.932, P =< 0.001
Binary thr: 0.41
Kappa(k): 0.728
C
TN
Patient's biopsy slide #2
Patient's biopsy slide #1
r: 0.939, P =< 0.001
Binary thr: 0.50
Kappa(k): 0.929
Binary predictions agreement
No
Yes

Figure 5.


A
HER2+
Highest predicted scores
Lowest predicted scores
B
ER+/HER2-
Highest predicted scores
Lowest predicted scores
C
TN
Highest predicted scores
Lowest predicted scores

Supplementary Figure S1.



Supplementary Table S1.

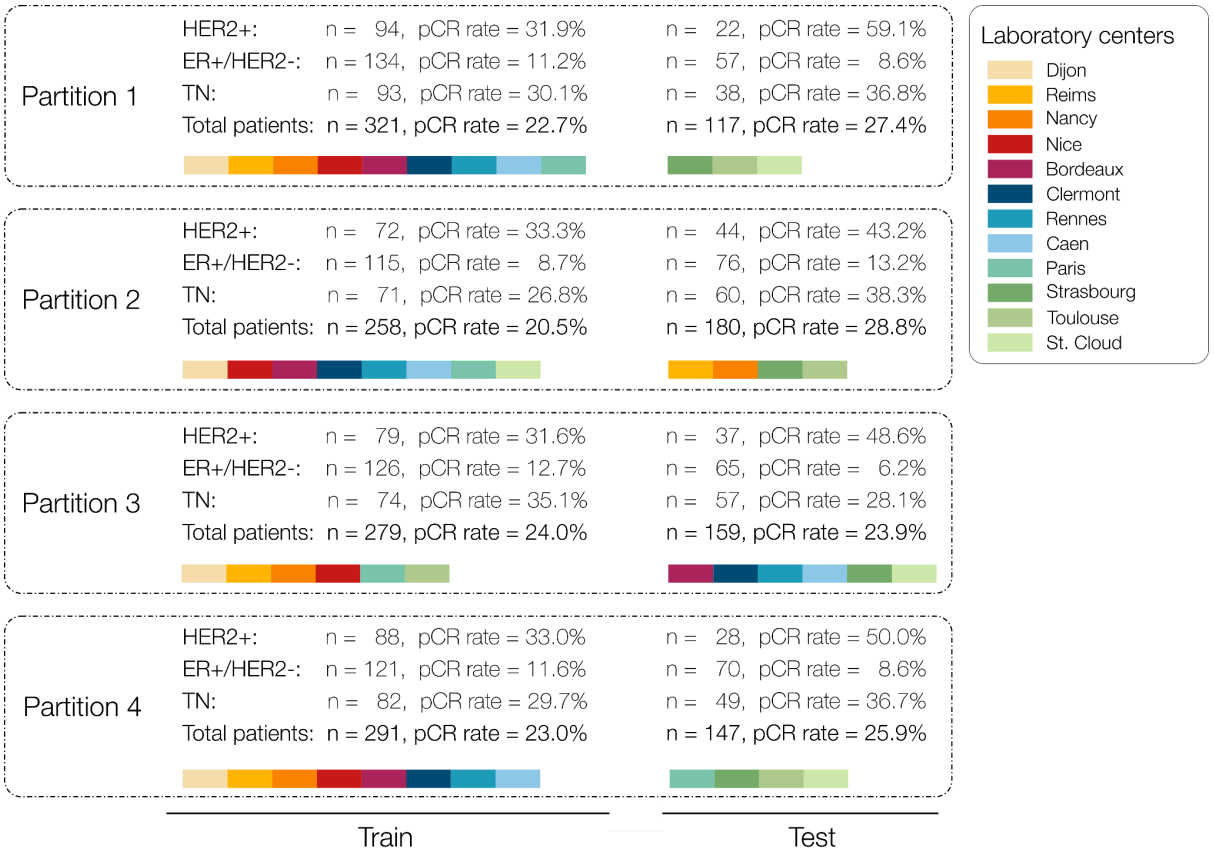
Center	Total patients	HER2+		ER+/HER2-		TN	
		pCR	RD	pCR	RD	pCR	RD
DIJON	112	13	23	5	48	6	17
STRASBOURG	69	10	3	4	24	9	19
REIMS	48	5	11	3	17	4	8
RENNES	23	2	4	0	6	2	9
NANCY	38	3	8	2	10	7	8
CAEN	34	3	6	0	14	3	8
NICE	26	2	4	4	8	2	6
PARIS	30	1	5	1	12	4	7

TOULOUSE	25	1	3	1	15	3	2
ST CLOUD	23	2	3	0	13	2	3
BORDEAUX	9	1	2	0	4	0	2
CLERMONT	1	0	1	0	0	0	0
TOTAL	**438**	**43**	**73**	**20**	**171**	**42**	**89**

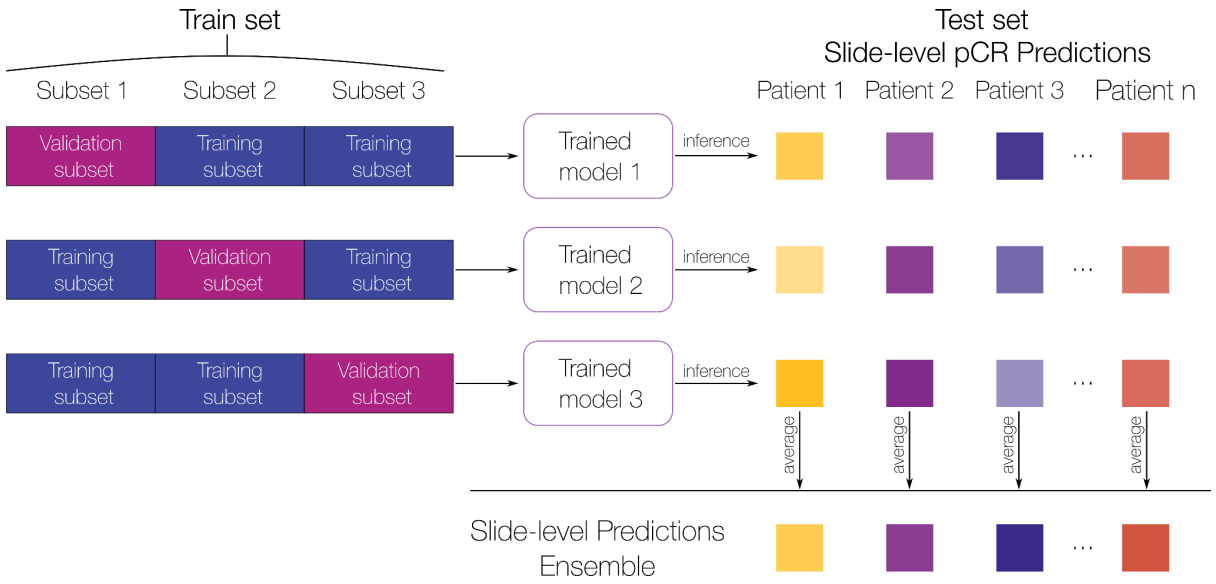
Supplementary Table S2.

Cancer subtype	Number of patients	pCR	RD	Prevalence
HER2+	220	47	173	0.214
ER+/HER2-	156	6	150	0.038
TN	114	35	79	0.307
Total	**490**	**88**	**402**	**0.180**

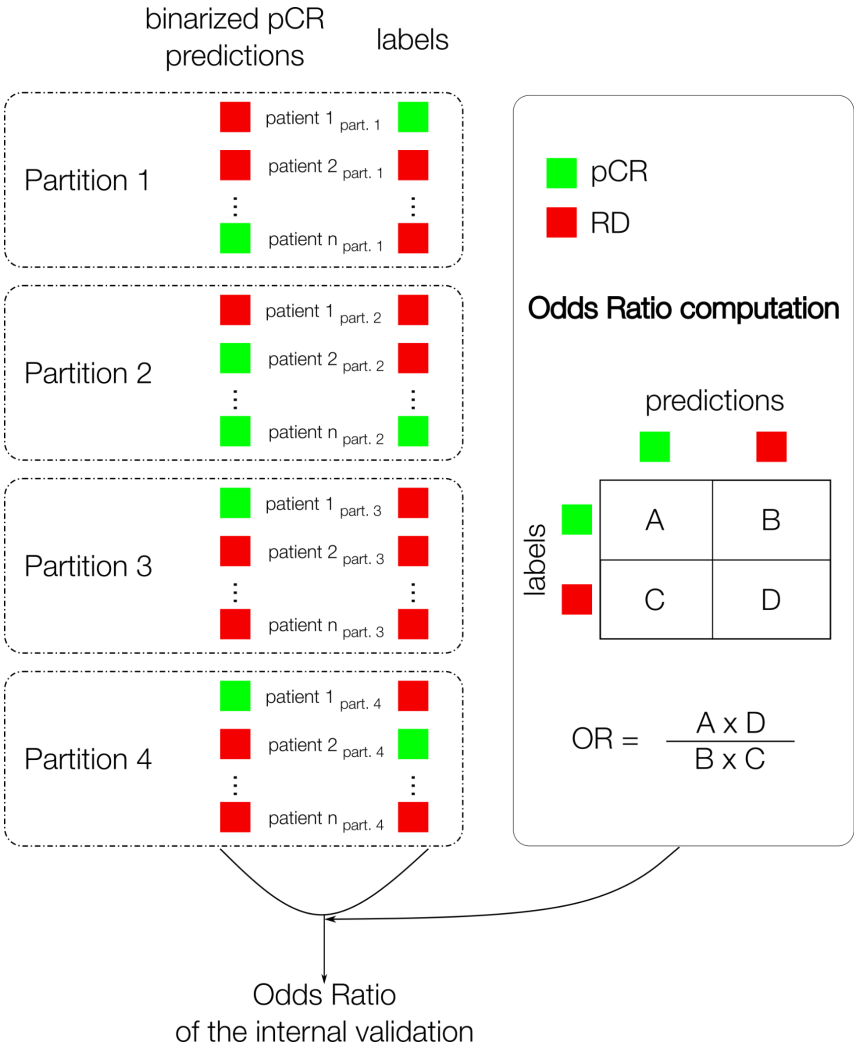
Supplementary Figure S2.



Supplementary Figure S3.



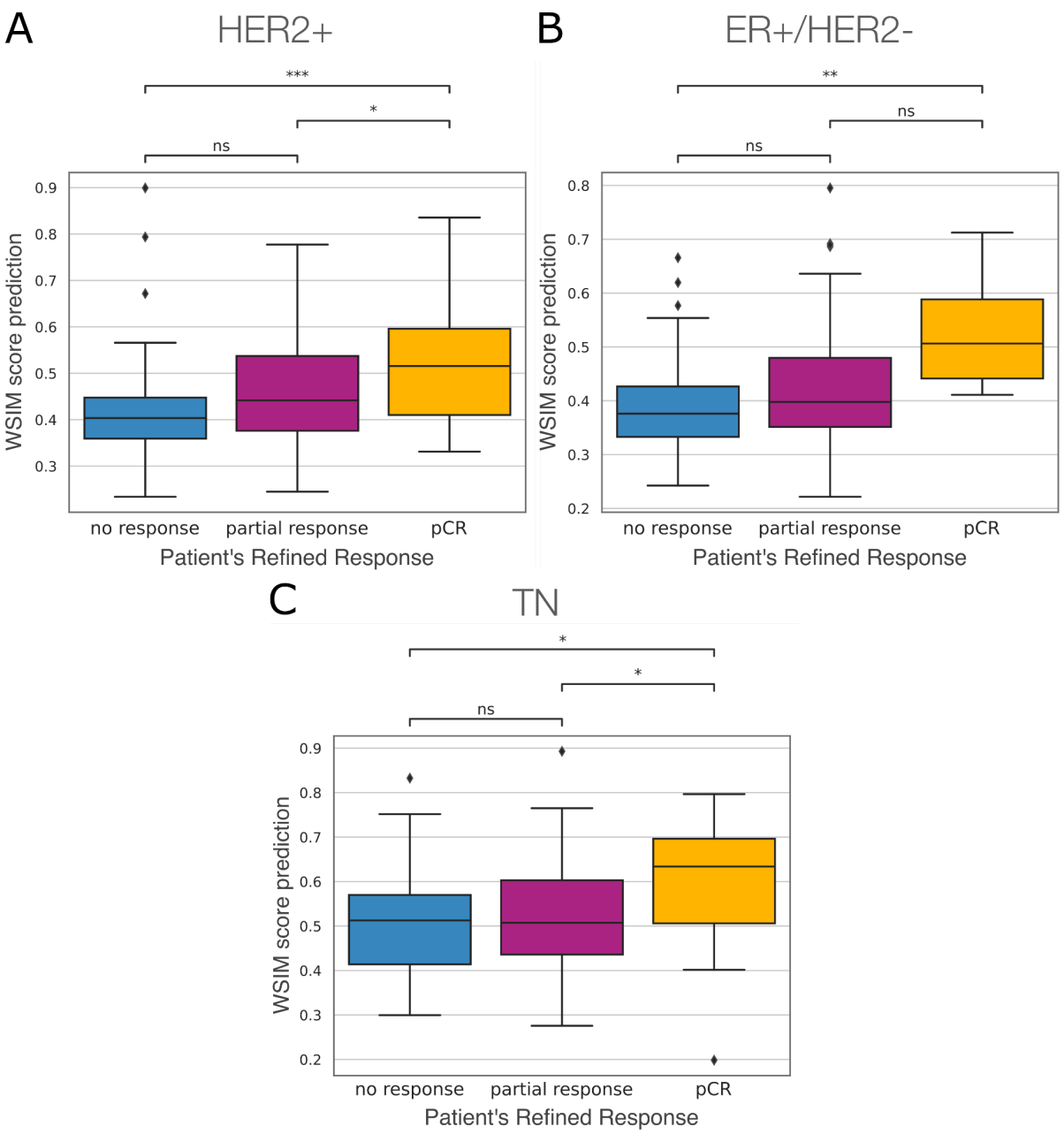
Supplementary Figure S4.



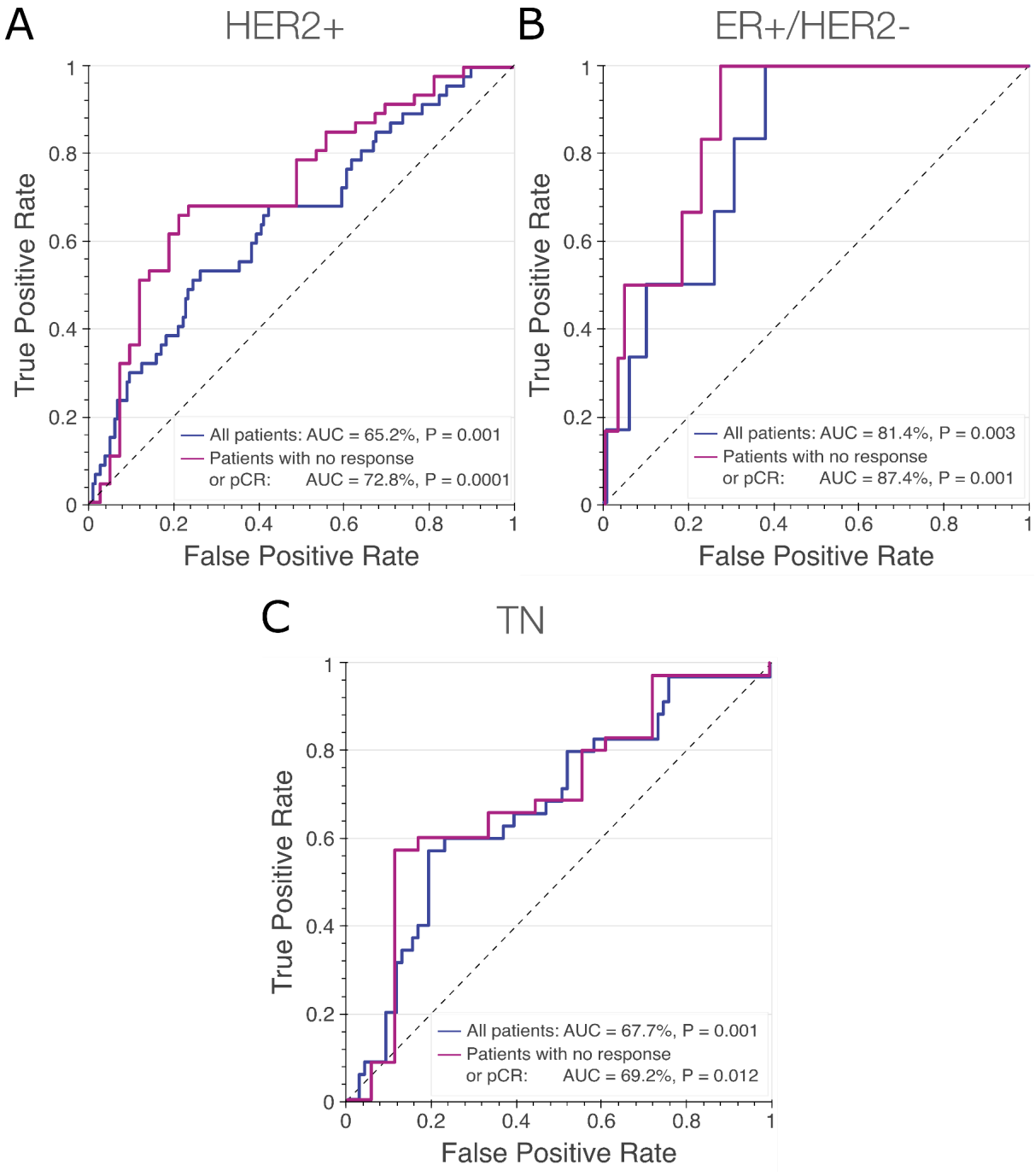
Supplementary Table S3.

Metrics	Sensitivity	Specificity	PPV	NPV
HER2+	0.872	0.283	0.248	0.891
ER+/HER2-	1	0.567	0.084	1
TN	0.800	0.430	0.383	0.829

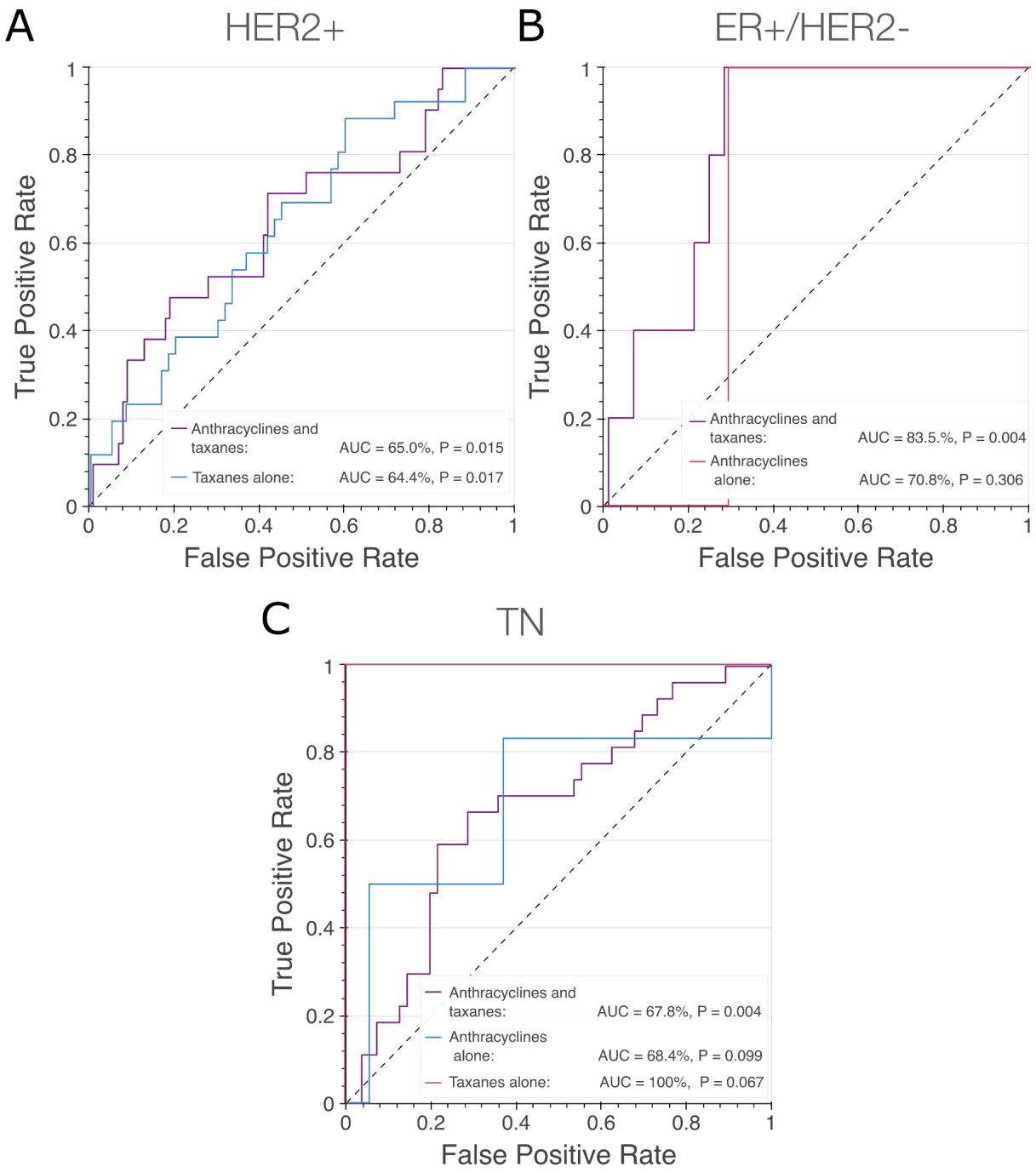
Supplementary Figure S5.



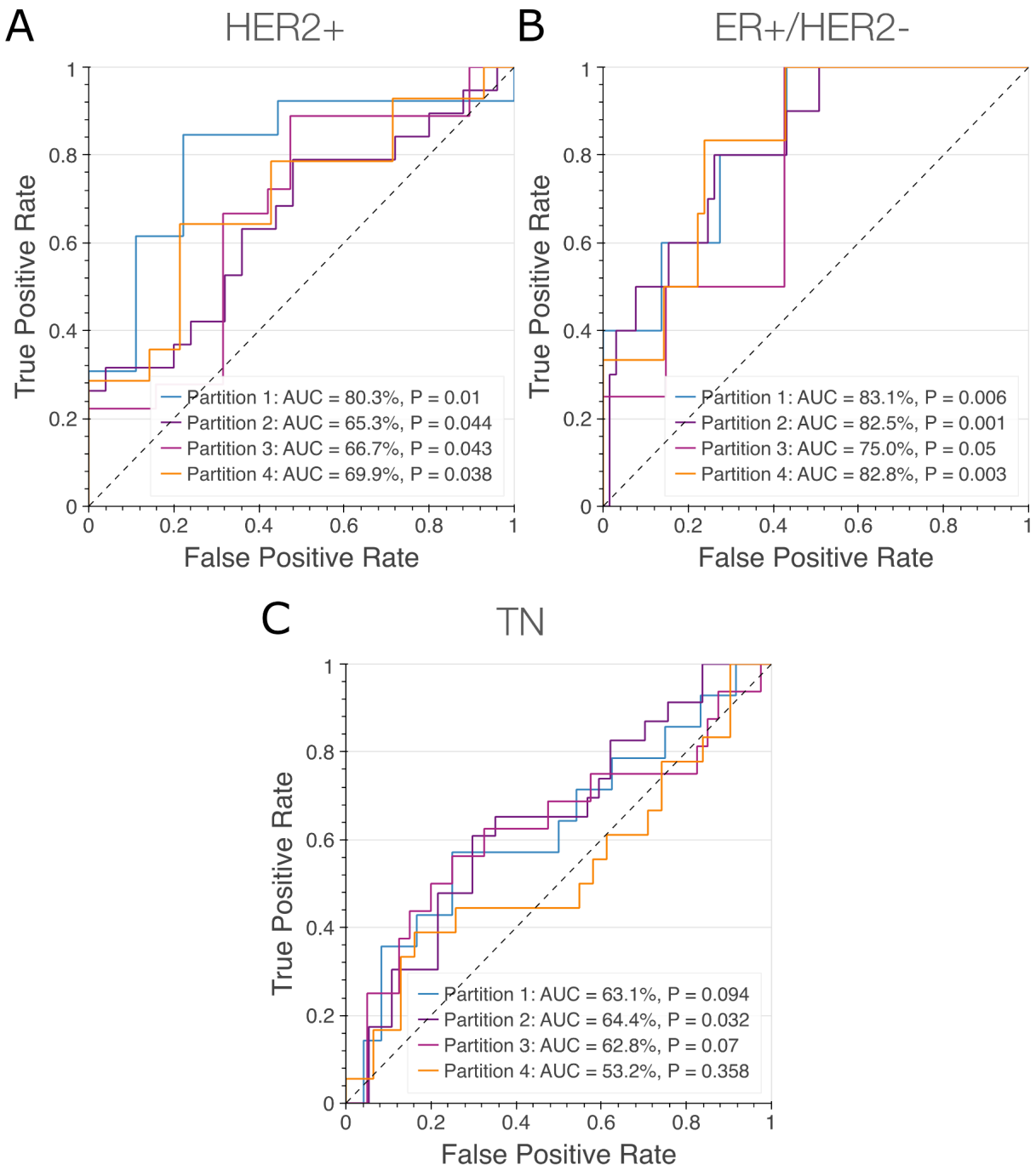
Supplementary Figure S6.



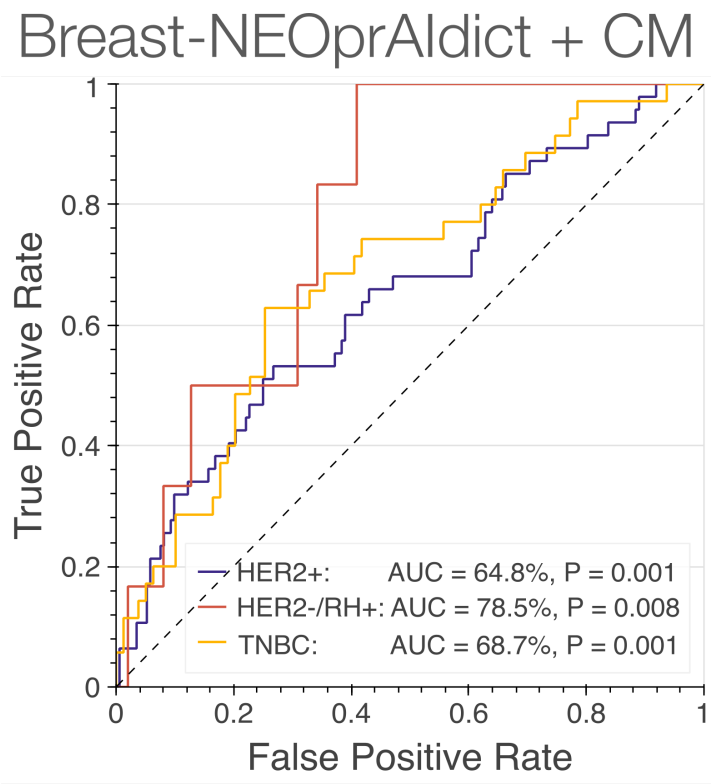
Supplementary Figure S7.



Supplementary Figure S8.



Supplementary Figure S9.



Supplementary Table S4.

Methods	EfficientNet B7-based model		ViT-S/16-based model		Breast-NEOprAIdict	
	AUC	***P***	**AUC**	***P***	**AUC**	***P***
HER2+	0.632	0.003	0.599	0.019	0.652	0.001
ER+/HER2-	0.776	0.01	0.746	0.02	0.814	0.003
TN	0.661	0.003	0.647	0.006	0.677	0.001