PLOS Computational Biology

RESEARCH ARTICLE

Predicting protein cascade expression from H&E images

Alejandro Leyva*, Abdul Rehman Akbar, Muhammad Khalid Khan Niazi

Department of Pathology, The Ohio State University, Columbus, Ohio, United States of America

* Leyva.29@osu.edu

Abstract

Protein expression within oncogenic or suppressive pathways is a hallmark indicator of oncogenesis. While traditional AI models in digital pathology attempt to predict singular proteins, there is a need to predict the downstream expression of proteins to indicate the propagation of signals. RNA expression provides novel information, but does not provide information about the downstream propagation of protein signals or whether those signals are functional. Using Reverse Phase Protein Array (RPPA) data with whole-slide images (WSIs) from the publicly available Cancer Genome Atlas Breast Adenocarcinoma dataset (TCGA-BRCA), we predict the expression of five key proteins identified from the apoptosis cascade, using DNA damage and repair (DDR) cascades as a biological control. Furthermore, we examine the performance of patch- level Vision Transformers (ViT) on the regression task, which was tested against the designed cellular-level ViT, CellRPPA. Our results demonstrate that patch-level vision transformers were unable to obtain statistically significant predictive results, achieving R-squared values < 0.1 for all folds. In addition, CellViT obtained R-squared values >0.1 in all five test folds. We also show that morphologically indicative cascades, such as the apoptosis cascade, provide significantly higher performance compared to the DDR cascade.

Author summary

We developed a method to estimate how groups of proteins behave inside tumors using standard tissue images that are routinely collected in clinical care. Rather than focusing on single molecules, we asked whether it is possible to capture broader biological processes such as how cells respond to stress or initiate programmed cell death directly from visual patterns in tissue structure. To do this, we trained a computational model to learn relationships between image features and protein measurements from the same tumors.

Check for updates

OPEN ACCESS

Citation: Leyva A, Akbar AR, Niazi MKK (2026) Predicting protein cascade expression from H&E images. PLoS Comput Biol 22(5): e1014262. https://doi.org/10.1371/journal.pcbi.1014262

Editor: Sophia Rudorf, Leibniz University Hanover, GERMANY

Received: January 30, 2026

Accepted: April 22, 2026

Published: May 4, 2026

Data availability statement: Datasets and diagnostic slides are publically available and anonymized: CellEcoNet gihub is available here: https://github.com/AI4Path-Lab/PathRosetta. CellRPPA and preprocessing code is available here: https://github.com/Alejandro21236/CellRPPA. Please see the zenodo link here: https://zenodo.org/records/19077131. Here

is the link to TCGABRCA, the openly available public dataset https://portal.gdc.cancer.gov/projects/ TCGA-BRCA.

Funding: The project described was supported in part by R01 CA276301 (PIs: MKKN and Dr. Wei Chen) from the National Cancer Institute, Pelatonia under IRP CC13702 (PIs: MKKN, Dr. Arya Mariam Roy, and Dr. Anna Vilgelm), The Ohio State University Department of Pathology and Comprehensive Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or National Institutes of Health or The Ohio State University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

We found that certain biological processes leave detectable signatures in tissue architecture, allowing the model to partially recover protein activity from images alone. However, this relationship is not complete, and much of the variation remains unexplained, highlighting the complexity of tumor biology.

Our work suggests that widely available pathology images may contain underused in- formation about underlying molecular processes. In the future, approaches like this could help provide additional biological insight in settings where direct molecular measurements are unavailable, supporting more informed research and potentially guiding clinical decision making.

1 Introduction

Breast adenocarcinoma is a highly documented disease and can be associated with a variety of genetic mutations that result in the propagation or suppression of protein signals [1]. Pathways can be intrinsically mediated, whereby the stimulus resulting in signal transduction is intracellular, or extrinsic. Within adenocarcinoma, common mutations in proteins such as TP53 result in the suppression of apoptotic signals by mutating the DNA-binding domain, resulting in dysfunctional proteins [2]. Other mutations, such as those in *PIK3CA* or *MYC*, result in the enrichment and propagation of the PIK3CT/AKT pathways by mutating kinase domains, leading to the overexpression of these proteins [3]. Apoptosis is responsible for regulating programmed cell death and is characterized by shrinkage, pyknosis, and reorganization of lipid structure [4]. The pathway is both intrinsically and extrinsically mediated by a combination of enzymes, receptors, and transcription factors [5]. Intrinsically mediated apoptosis occurs in two categorical fashions: negative, which is due to the absence of growth factors around a structure resulting in the triggering of cell death, or positive, which may be due to the presence of antigens, radiation, or hypoxia [6]. These changes result in the opening of inner mitochondrial pores due to lost membrane potential, preventing regular cell metabolism and causing the activation of BH3 proteins that detect metabolic stress [7]. As a result, BH3 proteins deactivate anti-apoptotic proteins such as BCL-2 and begin to activate apoptotic proteins such as BAK/BAX, resulting in the transduction of signals to XIAP (X-linked inhibitor of apoptosis protein), which then disinhibits the caspase family. The released caspases then catabolize the cell, resulting in cellular death [8].

The extrinsic pathway performs a similar function but is dependent on ligand–receptor binding of the TRAIL/FasL proteins, which results in the activation of caspases 8 and 10, which then activate the same caspases up to XIAP [9]. Common forms of extrinsic apoptosis include T-cell–mediated apoptosis and immune response [10].

The DNA damage and repair cascade is responsible for the repair of DNA in response to double-strand breaking or transcriptional errors. It has been demonstrated that lower

expression of the DDR cascade is prognostically predictive within breast adenocarcinoma and has been predictive of chemotherapy sensitivity and response [11]. DDR applications within clinical oncology include testing for mutations in genes responsible for double-strand breaking repair, *BRCA1* and *BRCA2* (Breast Cancer Gene 1/2), and can be secondarily characterized by RNA expression. However, the choice of genes to use for prognostic value is still debated [12]. Moreover, the expression of DDR genes is not physically visible via an electron microscope but can indirectly result in abnormal cellular growth or polyploid cells [13]. Genes that are often used to characterize the DDR cascade include *ATM, CHEK2, H2AFX, RAD51, and TP53*, among others. The ataxia–telangiectasia mutated gene (*ATM*) encodes the response to double-strand breaking and senses DNA double-strand breaks [14]. *RAD51* (radiation-sensitive protein 51) is responsible for the invasion of homologous DNA zones to allow for accurate and timely DNA repair and behaves as an ATPase [15]. *CHEK2* (checkpoint kinase 2) is a protein responsible for producing a kinase that cleaves DNA regions and is a radiation-sensitive protein that prevents the cell from entering mitosis upon DNA damage [16]. *CHEK2* has been shown to be a predictive biomarker for multiple organ cancers across Europe and North America [17]. *TP53* is a transcription factor responsible for encoding the p53 antigen, as well as a protein that activates or inhibits apoptotic genes such as *NOXA* [18]. *TP53* (tumor suppressor antigen 53) typically functions in multiple roles to suppress proliferative pathways and regulate cell cycling. *H2AFX* is responsible for regulating nucleosome formation and produces phosphorylative foci to recruit repair factors for double-strand breaks; the H2A family is also responsible for regulating chromatin accessibility and replication timing [19].

These cascades and their expression are typically predicted using RNA expression in the field of bioinformatics [20]. Within the field of digital pathology, gene expression models and molecular subtyping models have been developed using novel deep learning methods, whereby gene expression can be predicted from whole-slide image features [21]. More recently, multi- modal models integrating protein and genetic data are being developed, complementary to spatial transcriptomics and proteomics [22]. While RNA provides a perspective on cascade expression and prognosis, it does not indicate whether genes are translated and produced, or whether the proteins produced are functional [23]. Traditional models in the field have attempted to predict the expression of singular proteins or antigens that can be readily indi- cated from histology, including *HER2* expression and *EGFR* [24,25]; however, most models do not predict intracellular protein expression due to morphological ambiguity and poor generalization across cancers and external datasets. Multimodal models and comprehensive information on protein expression are required to understand protein functionalization for improved prognostic prediction. Information from proteomics contextualizes cellular behvav- ior and patterns in gene expression in systems biology, as cascades are highly interconnected and cross-talk mechanisms of action [26]. Misalignment between the extent of gene expression and the presence of proteins indicates translational or transcriptional dysfunction or deliberate inhibition.

Protein data are gathered from Reverse Phase Protein Array, which uses fluorescent antibodies that bind to the protein of interest [27]. Luminescence reflects antibody-based detection of target protein levels, while absorbance is used to quantify total protein loading for normalization. The sensitivity and accuracy of RPPA depend on the affinity, specificity, and availability of proteins to bind to the provided antibodies and are conditionally accurate. Typically, multimodal analyses between RNA and RPPA within public datasets are not performed due to the time differential at which each assay is performed, rendering direct comparison between protein and RNA expression invalid. When using AI for expression prediction, information is extracted from region/patch level of whole slide images [28]. In recent years, higher resolution image embeddings have been developed to examine Images at the cellular level, and are untested for protein prediction [29]. Since Protein Cascades are generally visible at the cellular level, there is a need to investigate the ability of AI at both the patch level and the cell level to predict the expression of multiple proteins.

In this study, we predict the expression of multiple intracellular proteins in aggregate as a regression task from WSIs. We present a comprehensive algorithm, CellRPPA, inspired by CellEcoNet [30], to compete with conventional patch-level algorithms. This study offers two novel investigations: i) differences between cell-level and patch-level

resolution in deep learning for protein expression prediction, and ii) the capacity for deep learning to predict morphologically ambiguous proteins, or proteins that are loosely indicated by histology [31]. As digital pathology continues to expand, there is an ever-growing need for multimodal models and proteomics to provide a comprehensive perspective on disease progression.

2 Materials and methods

From the publicly available Cancer Genome Atlas (TCGA) Breast Adenocarcinoma dataset (TCGA-BRCA), 919 RPPA samples were paired with whole-slide images. RPPA measure- ments were defined at the case level, yielding a 1:1 correspondence between each patient and a single protein cascade score. Multiple WSIs per case were included during training as independent inputs, and slide-level predictions were aggregated to obtain a single case- level estimate. Group-wise cross-validation ensured that all slides from a given patient were assigned to the same fold, preventing leakage. Cascades were defined using proteins. Data splitting was performed at the patient level to prevent leakage between training and test sets, and no samples were skipped or had missing information. The apoptosis cascade was defined as the expression of BCL2, BAX, XIAP, and cleaved caspases 3 and 7. This was done to represent the activation of both intrinsic pathways demonstrated by BCL and BAX, as well as the expression of the encompassing caspases and regulatory inhibitors. The DDR cascade was used as a morphologically ambiguous protein cascade to compare against the performance of apoptosis, which is microscopically noticeable. The proteins chosen to rep- resent the DDR cascade include H2AFX, CHEK2, TP53, TP53 BP1, and ATM; however, this is acknowledged to be a limited representation of the DDR cascade and was chosen to represent hallmark genes associated with prognostic value in breast adenocarcinoma.

The RPPA scores for the relative intensity of each protein were then summed and z-scored across each protein. If any protein that was part of the cascade was not included within a sample, the sample was excluded entirely. Cascade scores were defined as the sum of z- scored protein abundances to provide a compact proxy for coordinated pathway-associated protein expression. This formulation was not intended to model regulatory directionality or mechanistic pathway activity (e.g., opposing pro- and anti-apoptotic effects), but rather to capture aggregate multivariate protein signal as a supervised learning target. We acknowl- edge that this approach does not resolve antagonistic relationships among pathway members. A directionality-aware formulation, such as signed weighting or graph-based aggregation of protein interactions, represents an important future extension to more faithfully model reg- ulatory dynamics. The CellRPPA model is CellEcoNet reengineered with a regression head, and the GitHub repository for the model is included for both CellEcoNet and CellRPPA. Cell embeddings are derived from localized image regions, but CellRPPA is not simply a higher-resolution patch model; it integrates cell-level features within a structured hierarchi- cal aggregation that captures both local and contextual information. The patch-level ViT baseline already tests patch-only input under the same task and performs worse, indicating that the improvement is not explained by resolution alone. Cell embeddings were extracted using Trident and CellViT++, at 20 × magnification, and cell types were not determined.

The apoptosis and DDR tasks were performed separately, and training and evaluation took 150 hours for each task. The patch-level ViT uses a standard ViT-S/16 with cross-fold validation, where the DDR and apoptosis tasks were tested and validated separately and were not backpropagated jointly. We selected a standard ViT-S/16 baseline to isolate the effect of resolution and aggregation strategy rather than architectural optimization. Implementation parameters are shown in Table 1 for the S/16 vision transformer.

All parameterizations for both models are provided within anonymized shell scripts and should be able to run, provided that the necessary label files are available. Fig 1 shows the workflow for the study design, starting from the development of ground truth through the evaluation of each model. All computational experiments were performed on the Ohio Supercomputer on Nvidia A100 GPUs. All analyses were performed on the entire cohort, and the RPPA analytics were derived from the RPPA CSV using matplotlib, NumPy, and SciPy.

Table 1. Training configuration and hyper parameters used for regression.

Parameter	Value
Epochs	50
Number of folds	5
Batch size	1
Maximum patches per case	64
Learning rate	1×10^{-4}
Weight decay	0.05
Frozen transformer blocks	9
Random seed	42

https://doi.org/10.1371/journal.pcbi.1014262.t001

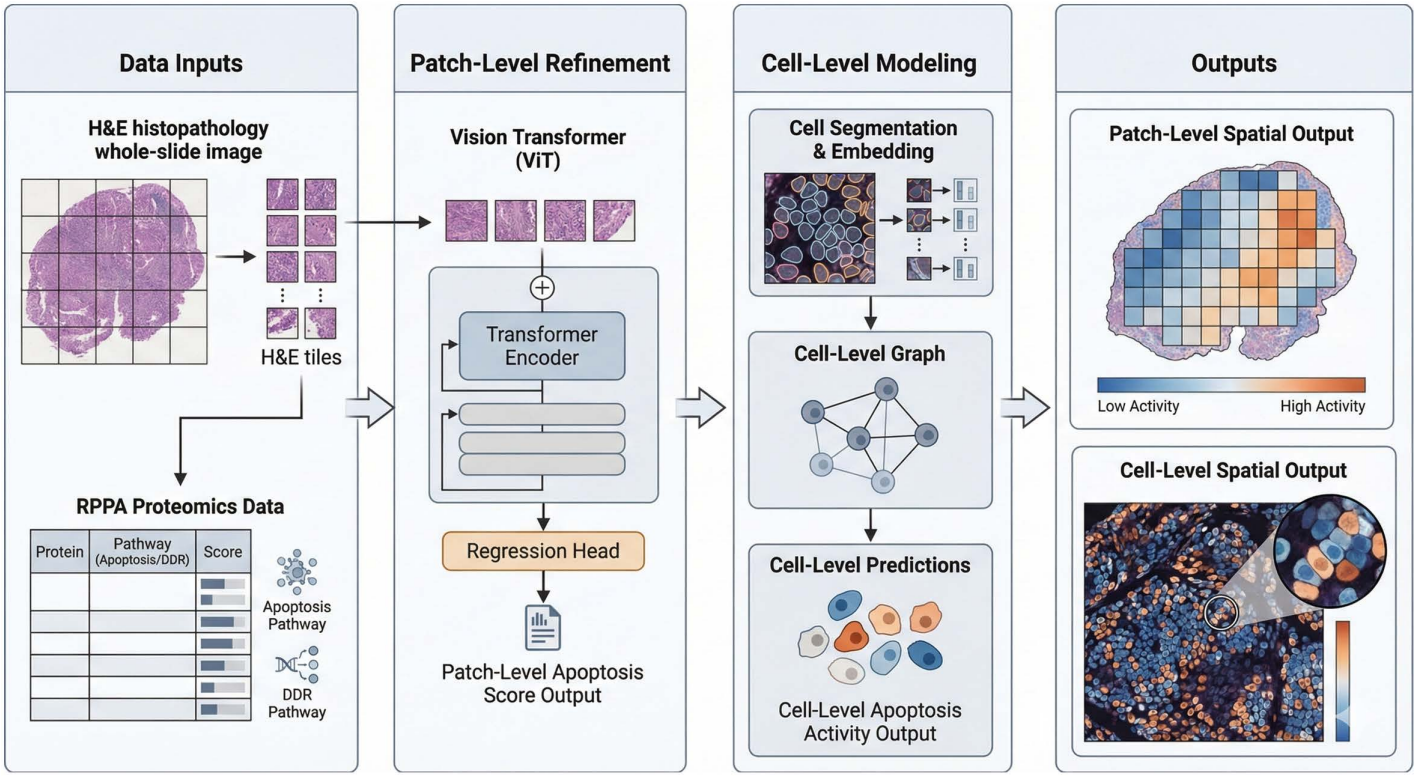

Fig 1. Overview of the cascade prediction framework. Whole-slide H&E images and RPPA-derived protein cascade scores serve as inputs. Patch-level modeling uses a Vision Transformer (ViT) to extract tile embeddings and regress pathway-level protein activity, while cell-level modeling performs cell segmentation, embedding, and graph construction to predict apoptosis activity at cellular resolution. The outputs illustrate spatially resolved pathway activity maps at both patch and cell scales, enabling comparison between coarse tissue-level predictions and fine-grained cellular cascade expression.

https://doi.org/10.1371/journal.pcbi.1014262.g001

3 Results

The labels for the protein scores were assigned by case, and each case was assigned into fold-wise training, testing, and validation cohorts for evaluation over 100 epochs per fold. Initial analysis presents results on the expression of proteins defined within each sample as aggregated scores, as shown in Fig 2. Fig 2A shows the correlation heatmap of protein abundance co-variation across each cascade. Proteins included within the apoptosis cascade exhibit moderate

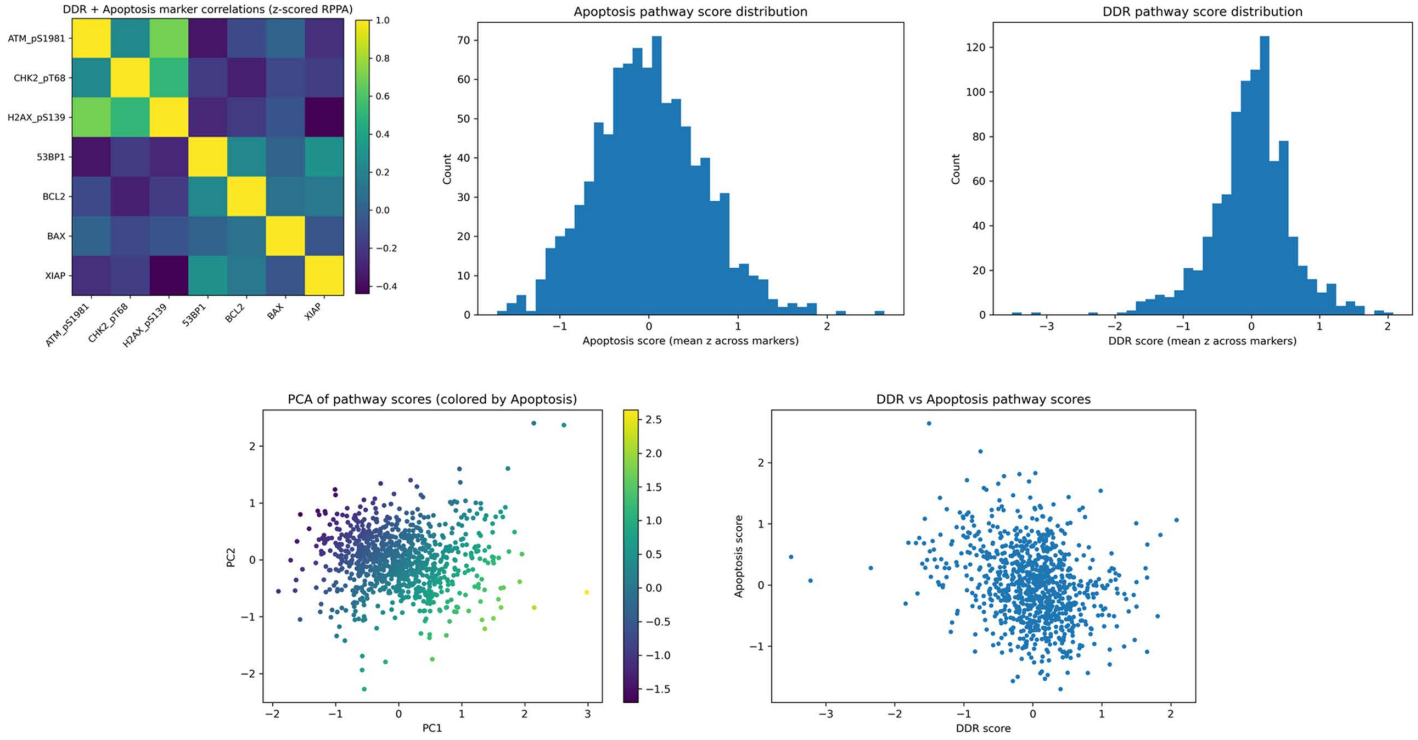

Fig 2. Proteomics and label-level analytics for apoptosis and DDR RPPA-derived path- way scores. Top row: inter-marker correlation structure and pathway score distributions. Bottom row: low-dimensional structure of pathway scores and cross-cascade correlation (control relationship between DDR and apoptosis). **(A)** Marker correlation heatmap (z-scored RPPA), (B) apoptosis score distribution, **(C)** DDR score distribution, **(D)** PCA of pathway scores colored by apoptosis, and **(E)** DDR versus apoptosis pathway scores.

https://doi.org/10.1371/journal.pcbi.1014262.g002

coordinated abundance among their counterpart proteins, consistent with shared cascade-level behavior. *BAX* and *XIAP* show very low co-variation, which is biolog- ically plausible given their opposing regulatory roles. Interestingly, *BCL2* and *TP53 BP1* show stronger abundance correlation, exceeding 50% Pearson correlation, while *TP53 BP1* shows negative abundance correlation with other DDR-associated proteins within its defined cascade. *TP53 BP1* also shows a strong correlation with *XIAP* abundance while exhibiting low co-variation with *BAX*. *BCL2* and *BAX* show low abundance correlation, which is consistent with the known biological roles of these proteins.

Within the DDR cascade, *TP53 BP1* shows lower co-variation with DDR-associated pro- teins, including *ATM*, *CHEK2*, and *H2AFX*. The abundance correlation between *ATM* and all other DDR proteins, except *TP53 BP1*, is remarkably strong, exceeding 80% with *H2AFX* and 50% with *CHEK2*. The correlation heatmap shows little overlap between DDR and apoptosis cascade abundance patterns and often demonstrates negative cross-cascade correlations. There is a small positive correlation between *BAX* and *ATM*, while most other cross-cascade correlations are below zero.

Fig 2B shows the standard Gaussian distribution of apoptosis pathway scores, with no observable skew and larger variance from the median. Fig 2C shows the DDR score distribution, which maintains a standard Gaussian distribution with outliers that are three standard deviations away from the mean, attributed to the absence of expression of any classified DDR gene. Fig 2D shows the PCA of pathway scores using UMAP projections for each sample, where samples that are closer together have similar protein expression profiles. A higher frequency of samples cluster toward minimal variance, while outliers, or cases with higher apoptosis cascade scores, are more dispersed and exhibit greater variance in protein expression, or markedly higher expression of proteins that are typically less expressed. To visualize the relationship between DDR scores and apoptosis, Fig 2E shows a plot that classifies samples by their respective scores. In general,

samples with higher apoptosis scores tend to have lower DDR scores, and vice versa. However, a large proportion of samples with apoptosis scores near zero also have DDR scores near zero, suggesting an inverse relationship.

The results for the patch-level ViT on the morphologically ambiguous DDR cascade, used as a control, are shown in Table 2. The model failed to obtain statistically significant correlations in three folds using False Discovery Rate (FDR) ($p < 0.05$) correction on the Spearman correlation for non-linear correlational analysis. Fold 2 completely failed to obtain any Pearson or Spearman correlation as a result of convergence, while the Mean Absolute Error (MAE) remained relatively high. Mean Squared Error (MSE) did not consistently decline across folds, and the MAE remained stagnant across all folds. The patch-level ViT is shown to predict close to no variance in protein expression within the DDR cohort across all five folds. Variability across folds likely reflects cohort heterogeneity and limited sample size. In addition, the Spearman and Pearson values were nearly identical in some folds, indicating similarity in morphological correlation.

The results for the patch-level ViT's prediction of apoptosis cascade expression are shown in Table 3. The results are noticeably worse for apoptosis, with only one of the five folds showing significant correlations and three of the five folds showing no correlation at all. The MAE also increased relative to the DDR cascade prediction, exceeding 0.5 in some folds, while the typical range of values is between 0 and 1. The MSE did not remain steady and did not consistently improve across folds; Fold 1 demonstrated the best MSE, which then steadily worsened. The model failed to explain any variance in protein expression and, in Fold 3, produced slightly negative values, demonstrating inadequate capability to predict apoptosis cascade expression, despite apoptosis being morphologically visible and the known correlation between protein expression and cascade expression.

The results for CellRPPA's prediction of the apoptosis cascade are shown in Table 4, demonstrating improvement over the patch-level ViT, which indicates that higher resolution can be advantageous for the task. For each fold, the epoch with the lowest validation loss was chosen to represent performance across folds. All folds explained at least 10% of the variance in the protein prediction task, while a substantial portion of folds in both the test and validation sets exceeded 20 or 30%. While the MAE remains consistent across cohorts, there is controlled variance across folds. The Pearson and Spearman correlation coefficients across all cohorts were above 40%, with the exception of one fold in the test set. The MSE remained consistent relative to MAE values across cohorts and folds, ranging between 29–35, with the exception of two instances in the test set.

Table 2. Performance metrics for DDR pathway prediction across folds. Erroneous values or failed convergences are reported as NaN for all tables.

Fold	Pearson *r*	Pearson *p*	Spearman *ρ*	Spearman *p*	MAE	MSE	R^2	FDR sig
1	0.1479	0.0870	0.1513	0.0799	0.4219	0.3009	0.0135	False
2	NaN	NaN	NaN	NaN	0.4437	0.4088	-0.0011	False
3	0.1243	0.1510	0.1803	0.0364	0.4286	0.3674	-0.0199	False
4	0.2507	0.0032	0.2507	0.0032	0.4106	0.3170	-0.0019	True
5	0.2169	0.0115	0.1418	0.1010	0.4156	0.3497	0.0320	True

https://doi.org/10.1371/journal.pcbi.1014262.t002

Table 3. Performance metrics for apoptosis pathway prediction across folds. Undefined correlation values are reported as NaN.

Fold	Pearson *r*	Pearson *p*	Spearman *ρ*	Spearman *p*	MAE	MSE	R^2	FDR sig
1	0.0392	0.6515	0.0445	0.6082	0.4184	0.2955	0.0003	False
2	NaN	NaN	NaN	NaN	0.5120	0.4067	-0.0116	False
3	NaN	NaN	NaN	NaN	0.4836	0.3702	-0.0330	False
4	0.2236	0.0089	0.2530	0.0030	0.5302	0.4424	0.0029	True
5	NaN	NaN	NaN	NaN	0.4610	0.3313	-0.0018	False

https://doi.org/10.1371/journal.pcbi.1014262.t003

Table 4. Cell-level apoptosis prediction performance across five-fold cross-validation. Re-ported metrics include Pearson correlation, Spearman correlation, mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2).

Split	Fold	Pearson	Spearman	MAE	MSE	R^2
	Fold 1	0.5734	0.5626	0.3811	0.2321	0.3257
	Fold 2	0.6147	0.6097	0.3846	0.3421	0.3773
Train	Fold 3	0.6009	0.5733	0.3808	0.2315	0.3572
	Fold 4	0.6072	0.5859	0.3906	0.2327	0.3926
	Fold 5	0.5969	0.5819	0.3737	0.2332	0.3515
	Fold 1	0.5000	0.4508	0.4771	0.3584	0.2446
	Fold 2	0.4300	0.4274	0.4144	0.2821	0.1708
Validation	Fold 3	0.4564	0.4579	0.4806	0.3551	0.2064
	Fold 4	0.4736	0.4669	0.4419	0.3083	0.1601
	Fold 5	0.5564	0.5754	0.4326	0.2915	0.3068
	Fold 1	0.4982	0.5199	0.4177	0.2800	0.2414
	Fold 2	0.3770	0.3574	0.4645	0.2332	0.1222
Test	Fold 3	0.4629	0.4688	0.4227	0.2855	0.1659
	Fold 4	0.4454	0.4439	0.4469	0.3135	0.1648
	Fold 5	0.5084	0.4755	0.4220	0.2783	0.2484

https://doi.org/10.1371/journal.pcbi.1014262.t004

Table 5 shows the prediction performance for the DDR cascade using CellRPPA, which demonstrated worse values in comparison to apoptosis prediction performance. Relative to the patch-level tasks, CellRPPA showed no improvement over the patch-level ViT in terms of explained variance. Similar to the patch-level ViT, the Spearman and Pearson correlation coefficients fall within the same range, and the PCC values are similar, if not equal, to the Spearman coefficients. The MAE remained within a consistent range, as observed in the other predictive tasks, and was relatively stable across folds but did not exhibit a noticeable decrease.

Table 5. Cell-level DNA damage response (DDR) prediction performance across five-fold cross-validation. Metrics include Pearson correlation, Spearman correlation, mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2).

Split	Fold	Pearson	Spearman	MAE	MSE	R^2
	Fold 1	0.3200	0.2700	0.4000	0.3110	0.1000
	Fold 2	0.3600	0.3600	0.3900	0.2900	0.1300
Train	Fold 3	0.2700	0.2600	0.4100	0.3400	0.0500
	Fold 4	0.3000	0.3000	0.4200	0.3400	0.0900
	Fold 5	0.0900	0.0500	0.4100	0.3300	0.0000
	Fold 1	0.2400	0.2400	0.4500	0.3540	0.0400
	Fold 2	0.1500	0.1000	0.4500	0.3500	-0.0400
Validation	Fold 3	0.0900	0.1100	0.4100	0.2700	-0.0500
	Fold 4	0.2200	0.2200	0.4300	0.3100	-0.1090
	Fold 5	0.1800	0.2400	0.4000	0.3000	0.0000
	Fold 1	0.1700	0.2000	0.4100	0.3200	0.0000
	Fold 2	0.2200	0.2200	0.4100	0.3500	0.0000
Test	Fold 3	0.1300	0.1200	0.4400	0.3500	-0.0500
	Fold 4	0.0000	-0.0200	0.4500	0.3700	-0.2500
	Fold 5	0.1300	0.1400	0.4800	0.4400	-0.0500

https://doi.org/10.1371/journal.pcbi.1014262.t005

The results demonstrate the failure of patch-level ViTs on multi-protein prediction for apoptosis and DDR. In contrast, for cell-level ViTs (CellRPPA), the model succeeds in pre-

dicting apoptosis but fails on DDR, reflecting the histological ambiguity of these proteins. Validation of protein choices for both DDR and apoptosis demonstrates that the selected proteins behaved within the same cascade and function and also exhibited noticeable co- expression. It is also noted that samples with lower DDR expression had higher apoptosis protein expression as measured by RPPA. Across folds, CellRPPA demonstrated improved predictive performance for apoptosis ($R2 = 0.189 \pm 0.054$) compared to the patch-level ViT ($R2 = -0.009 \pm 0.015$). In contrast, both models failed to explain variance in the DDR task (patch-level: $R2 = 0.005 \pm 0.019$; CellRPPA: $R2 = -0.070 \pm 0.105$), consistent with the morphological ambiguity of DDR-associated proteins.

4 Discussion

Analysis of breast adenocarcinoma suggests that, in general, there is an inverse relationship between DDR expression and apoptosis expression, which is validated by the correlation heatmap shown in Fig 2A, indicating that a functional basis exists within breast adeno- carcinoma. Previous studies indicate that the apoptosis and DNA damage response cascades are tightly coupled rather than independent. DDR signaling functions upstream to assess genomic damage and, when repair fails, promotes apoptotic commitment. Following irre- versible activation of executioner caspases, key DDR components are cleaved, effectively terminating DNA repair processes [32]. While a few outliers exist in each cohort, the analysis suggests that most proteins exhibit similar levels of expression within the cohort. We acknowledge that this formulation does not account for opposing regulatory roles within pathways. Future work will incorporate directionality-aware weighting or graph-based ag- gregation to better reflect pathway dynamics.

Apoptosis is well known to be visible under compound or fluorescent microscopy, and it is notable that the patch-level ViT fails to detect and predict cascade expression and performs markedly worse than on the DDR cascade. Performance on the DDR cascade in both trials demonstrates that prediction is not suitable at either resolution, likely due to the inherent morphological ambiguity of DDR expression in histology. DDR expression may reflect replicative stress, aneuploidy, and other phenomena that can manifest in multiple visible forms across different cell types. In contrast, apoptosis has been observed to exhibit a set of well-characterized behaviors that are similar across most cell types. It should be noted that patch-level ViT does not provide sufficient resolution to statistically learn the patterns that characterize apoptosis. These results suggest that cascade prediction is better for proteins with histologically visible behaviors and that higher-resolution representations can be advantageous for such cascades. Future work will extend this model to additional pathways (e.g., Reactome) to assess pathway-specific predictability.

Since cellular models are proven to be capable of predicting aggregate protein activity, it makes sense to move toward modeling co-regulation and interactions within protein cas- cades in a spatially resolved way for better explainability and predictive power. With the rise of spatial transcriptomics in digital pathology, interactions between spatial proteomics and transcriptomics can now be modeled at the cellular level using platforms like Xenium and Orion without temporal mismatch between measurements [33,34]. Prior work has explored gene expression prediction using graph-based neural networks, where gene–gene interactions are modeled through distance-aware edges and weighted connectivity [35]. More importantly, transcriptomics has been used to predict proteomics through deep learning, and given the underlying biological coupling, the inverse direction, predicting transcriptomic states from proteomic signals, is also a reasonable extension [36]. Other approaches include contrastive learning across spatial transcriptomic images to capture differences in tumor microenvironments and cellular composition, enabling gene expression inference from contextual variation [37]. Cellular-level models provide the resolution needed to compare microenvironments within the same sample while also capturing transcriptomic and proteomic interactions locally. As a practical objective, cascade-derived features can be used for drug response prediction and survival modeling, as shown by Reitsam et al., allowing AI systems to capture underlying biological dynamics rather than just static measurements [38]. These cascade features can also be extended to tasks like lymph node metastasis prediction or patient risk stratification based on aggregate expression

patterns [39]. That said, there is still a gap in integrating full cascade-level information across modalities. While some work has imputed metabolite profiles from RNA, true integration across metabolomics, proteomics, and transcriptomics remains limited [40,41]. The core idea here is that histology can act as the anchor for integrating these signals, enabling cellular-level proteomic cascade prediction directly from tissue phenotype. But realistically, modeling cascades from phenotype back to genotype in digital pathology is still wide open and far from solved.

This study focuses on establishing proof-of-concept within a controlled dataset. External validation across institutions remains an important direction for future work. Given the known domain shift in histopathology, cross-cohort generalization is non-trivial and warrants dedicated study.

5 Limitations

While this study uses the term "cascades" to describe multi-protein prediction, each cascade represents a characterized set of hallmark proteins from canonically observed pathways. Full representation of each cascade would require a substantially larger protein cohort, which would introduce additional noise and increase prediction difficulty. We also note that the model choices used for the patch-level ViT are limited and not fully generalizable, and similar limitations apply to CellViTs. Future studies should derive more generalizable conclusions across omics prediction by conducting broader evaluations of currently available models to assess the advantages of cell-level ViTs on omics tasks. In this study, minimal architectures were used to establish baseline performance for multiprotein prediction.Future work will include comparisons to MIL-based and pathology-optimized architectures.

6 Conclusion

Using cell-level ViTs and patch-level ViTs, we predict the expression of hallmark proteins within cascades and assess the advantages and disadvantages of each modeling approach. Our results show that cascade prediction is a viable task and can provide insight into molecular, genetic, and cellular functions and responses. Overall, we consider the performance of CellRPPA modest but promising in comparison to other omics-prediction tasks, though improvement in MAE is needed for viability. We demonstrate that patch-level ViTs do not outperform CellRPPA in microscopically visible cascades and that CellRPPA does not provide an advantage over patch-level models for morphologically ambiguous proteins within the DDR cascade. It can be concluded that CellRPPA does not provide an advantage over patch-level ViTs for cascade prediction when targeting morphologically ambiguous proteins.

Future studies should extend this analysis to additional cascades to better understand the integration of deep learning–derived proteomics into clinical informatics. In addition, cross-cohort analyses are needed to demonstrate generalization across cancer types.

Author contributions

Conceptualization: Alejandro Leyva.

Data curation: Abdul Rehman Akbar.

Formal analysis: Alejandro Leyva.

Funding acquisition: Muhammad Khalid Khan Niazi.

Investigation: Alejandro Leyva.

Methodology: Alejandro Leyva.

Project administration: Muhammad Khalid Khan Niazi.

Software: Abdul Rehman Akbar.

Supervision: Muhammad Khalid Khan Niazi.

Validation: Alejandro Leyva.

Visualization: Alejandro Leyva.

Writing – original draft: Alejandro Leyva.

Writing – review & editing: Muhammad Khalid Khan Niazi.

References

1. Tomlinson IANPM. Mutations in normal breast tissue and breast tumours. Breast Cancer Res. 2001;3(5). https://doi.org/10.1186/bcr311
2. Shahbandi A, Nguyen HD, Jackson JG. TP53 mutations and outcomes in breast cancer: reading beyond the headlines. Trends Cancer. 2020;6(2):98–110. https://doi.org/10.1016/j.trecan.2020.01.007 PMID: 32061310
3. Jenkins ML, Ranga-Prasad H, Parson MAH, Harris NJ, Rathinaswamy MK, Burke JE. Oncogenic mutations of PIK3CA lead to increased membrane recruitment driven by reorientation of the ABD, p85 and C-terminus. Nat Commun. 2023;14(1):181. https://doi.org/10.1038/s41467-023-35789-6 PMID: 36635288
4. Elmore S. Apoptosis: a review of programmed cell death. Toxicol Pathol. 2007;35(4):495–516. https://doi.org/10.1080/01926230701320337
5. Lossi L. The concept of intrinsic versus extrinsic apoptosis. Biochem J. 2022;479(3):357–84. https://doi.org/10.1042/BCJ20210854
6. Xu G, Shi Y. Apoptosis signaling pathways and lymphocyte homeostasis. Cell Research. 2007;17(9):759–71. https://doi.org/10.1038/cr.2007.52
7. Lomonosova E, Chinnadurai G. BH3-only proteins in apoptosis and beyond: an overview. Oncogene. 2008;27 Suppl 1(Suppl 1):S2-19. https://doi.org/10.1038/onc.2009.39 PMID: 19641503
8. Chaudhary AK. A potential role of x-linked inhibitor of apoptosis protein in mitochondrial membrane permeabilization and its implication in cancer therapy. Drug Discov Today. 2016;21(1):38–47. https://doi.org/10.1016/j.drudis.2015.07.014
9. Knight MJ, Riffkin CD, Muscat AM, Ashley DM, Hawkins CJ. Analysis of FasL and TRAIL induced apoptosis pathways in glioma cells. Oncogene. 2001;20(41):5789–98. https://doi.org/10.1038/sj.onc.1204810 PMID: 11593384
10. Grafman J, Salazar AM. Traumatic brain injury. Elsevier; 2015.
11. Li W, et al. Implications of DNA damage response and immunotherapy in tumor therapy. Cell Commun Signal. 2025;23(1). https://doi.org/10.1186/s12964-025-02527-y
12. Abu-Helalah M, et al. BRCA1 and BRCA2 genes mutations among high risk breast cancer patients in Jordan. Scientific Reports. 2020;10(1):17573. https://doi.org/10.1038/s41598-020-74250-2
13. Zheng L, Dai H, Zhou M, Li X, Liu C, Guo Z, et al. Polyploid cells rewire DNA damage response networks to overcome replication stress-induced barriers for tumour progression. Nat Commun. 2012;3:815. https://doi.org/10.1038/ncomms1825 PMID: 22569363
14. Ueno S, Sudo T, Hirasawa A. ATM: functions of ATM kinase and its relevance to hereditary tumors. Int J Mol Sci. 2022;23(1):523. https://doi.org/10.3390/ijms23010523 PMID: 35008949
15. Meyer D, et al. Rad51 determines pathway usage in post-replication repair. Nature Commun. 2026. https://doi.org/10.1038/s41467-025-68109-1
16. Vahteristo P, Bartkova J, Eerola H, Syrjäkoski K, Ojala S, Kilpivaara O, et al. A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. Am J Hum Genet. 2002;71(2):432–8. https://doi.org/10.1086/341943 PMID: 12094328
17. Cybulski C, Górski B, Huzarski T, Masojć B, Mierzejewski M, Debniak T, et al. CHEK2 is a multiorgan cancer susceptibility gene. Am J Hum Genet. 2004;75(6):1131–5. https://doi.org/10.1086/426403 PMID: 15492928
18. Shibue T, Takeda K, Oda E, Tanaka H, Murasawa H, Takaoka A, et al. Integral role of Noxa in p53-mediated apoptotic response. Genes Dev. 2003;17(18):2233–8. https://doi.org/10.1101/gad.1103603 PMID: 12952892
19. Yin X, Zeng D, Liao Y, Tang C, Li Y. The function of H2A histone variants and their roles in diseases. Biomolecules. 2024;14(8):993. https://doi.org/10.3390/biom14080993 PMID: 39199381
20. Henninger JE, Young RA. An RNA-centric view of transcription and genome organization. Mol Cell. 2024;84(19):3627–43. https://doi.org/10.1016/j.molcel.2024.08.021
21. Pizurica M, Zheng Y, Carrillo-Perez F, Noor H, Yao W, Wohlfart C, et al. Digital profiling of gene expression from histology images with linearized attention. Nat Commun. 2024;15(1):9886. https://doi.org/10.1038/s41467-024-54182-5 PMID: 39543087
22. Li Z, Li Y, Xiang J, Wang X, Yang S, Zhang X, et al. AI-enabled virtual spatial proteomics from histopathology for interpretable biomarker discovery in lung cancer. Nat Med. 2026;32(1):231–44. https://doi.org/10.1038/s41591-025-04060-4 PMID: 41491099
23. Moore JB, Weeks ME. Proteomics and systems biology: current and future applications in the nutritional sciences. Adv Nutri. 2011;2(4):355–64. https://doi.org/10.3945/an.111.000554
24. Jiao P. Prediction of HER2 status based on deep learning in H&E-stained histopathology images of bladder cancer. Biomed. 2024;12(7). https://doi.org/10.3390/biomedicines12071583
25. Park J, Shin S, Hwang W, Keum S, Brattoli B, Rawson JH, et al. Deep learning predicts EGFR mutation status from histology images in non-small cell lung cancer. Cancer Res Commun. 2025;5(12):2127–41. https://doi.org/10.1158/2767-9764.CRC-25-0155 PMID: 41211715

26. Nadendla EK, Tweedell RE, Kasof G, Kanneganti T-D. Caspases: structural and molecular mechanisms and functions in cell death, innate immunity, and disease. Cell Discov. 2025;11(1):42. https://doi.org/10.1038/s41421-025-00791-3 PMID: 40325022
27. Coarfa C, et al. Reverse-phase protein array: technology, application, data processing, and integration. J Biomol Tech. 2021;32(1):15–29. https://doi.org/10.7171/jbt.21-3202-001
28. Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. 2020. https://doi.org/arXiv:2010.11929
29. Hörst F. CellViT: vision transformers for precise cell segmentation and classification. 2023. https://doi.org/arXiv:2306.15350
30. Akbar ARE. CellEcoNet: decoding the cellular language of pathology with deep learning for invasive lung adenocarcinoma recurrence prediction. arXiv preprint. 2025. https://arxiv.org/abs/2508.16742
31. Wu T. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. The Innovation. 2021;2(3):100141. https://doi.org/10.1016/j.xinn.2021.100141
32. De Zio D, Cianfanelli V, Cecconi F. New insights into the link between DNA damage and apoptosis. Antioxid Redox Signal. 2013;19(6):559–71. https://doi.org/10.1089/ars.2012.4938 PMID: 23025416
33. Yin M, et al. DNA damage response and cancer metastasis: clinical implications and therapeutic opportunities. Exon Publications; 2022.
34. 10x Genomics. Visium spatial gene expression. 2020. https://www.10xgenomics.com/products/visium-spatial-gene-expression
35. RareCyte, Inc. Orion spatial biology platform. RareCyte technical documentation. 2023. https://rarecyte.com/orion/
36. Li B. Gene expression prediction from histology images via hypergraph neural networks. Brief Bioinform. 2024;25(6). https://doi.org/10.1093/bib/bbae500
37. Cranney CW, Meyer JG. Multi-dataset integration and residual connections improve proteome prediction from transcriptomes using deep learning. bioRxiv. 2024. https://doi.org/10.1101/2024.07.08.602560
38. Wang Q, Chen W-J, Su J, Wang G, Song Q. HECLIP: histology-enhanced contrastive learning for imputation of transcriptomics profiles. Bioinformatics. 2025;41(7):btaf363. https://doi.org/10.1093/bioinformatics/btaf363 PMID: 40569046
39. Reitsam NG, Enke JS, Vu Trung K, Märkl B, Kather JN. Artificial intelligence in colorectal cancer: from patient screening over tailoring treatment decisions to identification of novel biomarkers. Digestion. 2024;105(5):331–44. https://doi.org/10.1159/000539678 PMID: 38865982
40. Reitsam NG, Jiang X, Liang J, Grosser B, Grozdanov V, Loeffler CM, et al. Deep learning-based H&E-derived risk scores in colorectal cancer: associations with tumour morphology, biology, and predicted drug response. J Pathol. 2026;269(1):112–24. https://doi.org/10.1002/path.70039 PMID: 41716034
41. Xie AX, Tansey W, Reznik E. UnitedMet harnesses RNA-metabolite covariation to impute metabolite levels in clinical samples. Nat Cancer. 2025;6(5):892–906. https://doi.org/10.1038/s43018-025-00943-0 PMID: 40251399