Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: a retrospective multi-cohort study

Kexin Ding, Mu Zhou*, He Wang, Shaoting Zhang†, Dimitri N Metaxas†*

Summary

Background Digital whole-slide images are a unique way to assess the spatial context of the cancer microenvironment. Exploring these spatial characteristics will enable us to better identify cross-level molecular markers that could deepen our understanding of cancer biology and related patient outcomes.

Methods We proposed a graph neural network approach that emphasises spatialisation of tumour tiles towards a comprehensive evaluation of predicting cross-level molecular profiles of genetic mutations, copy number alterations, and functional protein expressions from whole-slide images. We introduced a transformation strategy that converts whole-slide image scans into graph-structured data to address the spatial heterogeneity of colon cancer. We developed and assessed the performance of the model on The Cancer Genome Atlas colon adenocarcinoma (TCGA-COAD) and validated it on two external datasets (ie, The Cancer Genome Atlas rectum adenocarcinoma [TCGA-READ] and Clinical Proteomic Tumor Analysis Consortium colon adenocarcinoma [CPTAC-COAD]). We also predicted microsatellite instability and result interpretability.

Findings The model was developed on 459 colon tumour whole-slide images from TCGA-COAD, and externally validated on 165 rectum tumour whole-slide images from TCGA-READ and 161 colon tumour whole-slide images from CPTAC-COAD. For TCGA cohorts, our method accurately predicted the molecular classes of the gene mutations (area under the curve [AUCs] from 82·54 [95% CI 77·41–87·14] to 87·08 [83·28–90·82] on TCGA-COAD, and AUCs from 70·46 [61·37–79·61] to 81·80 [72·20–89·70] on TCGA-READ), along with genes with copy number alterations (AUCs from 81·98 [73·34–89·68] to 90·55 [86·02–94·89] on TCGA-COAD, and AUCs from 62·05 [48·94–73·46] to 76·48 [64·78–86·71] on TCGA-READ), microsatellite instability (MSI) status classification (AUC 83·92 [77·41–87·59] on TCGA-COAD, and AUC 61·28 [53·28–67·93] on TCGA-READ), and protein expressions (AUCs from 85·57 [81·16–89·44] to 89·64 [86·29–93·19] on TCGA-COAD, and AUCs from 51·77 [42·53–61·83] to 59·79 [50·79–68·57] on TCGA-READ). For the CPTAC-COAD cohort, our model predicted a panel of gene mutations with AUC values from 63·74 (95% CI 52·92–75·37) to 82·90 (73·69–90·71), genes with copy number alterations with AUC values from 62·39 (51·37–73·76) to 86·08 (79·67–91·74), and MSI status prediction with AUC value of 73·15 (63·21–83·13).

Interpretation We showed that spatially connected graph models enable molecular profile predictions in colon cancer and are generalised to rectum cancer. After further validation, our method could be used to infer the prognostic value of multiscale molecular biomarkers and identify targeted therapies for patients with colon cancer.

Funding This research has been partially funded by ARO MURI 805491, NSF IIS-1793883, NSF CNS-1747778, NSF IIS 1763523, DOD-ARO ACC-W911NF, and NSF OIA-2040638 to Dimitri N Metaxas.

Lancet Digit Health 2022; 4: e787–95

See **Comment** page e766

*Contributed equally

†Joint last authors

Department of Computer Science, The University of North Carolina at Charlotte, Charlotte, NC, USA (K Ding MS); **Sensebrain Research, San Jose, CA, USA** (M Zhou PhD); **Department of Pathology, Medical School, Yale University, New Haven, CT, USA** (H Wang MD); **Shanghai Artificial Intelligence Laboratory and SenseTime Research, Shanghai, China** (S Zhang PhD); **Department of Computer Science, Rutgers University, Piscataway, NJ, USA** (Prof D N Metaxas PhD)

Correspondence to:
Prof Dimitri N Metaxas, Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA
dnm@cs.rutgers.edu

Introduction

Colon cancer is the third most common cancer worldwide and accounted for 10% of all new detected cancers and approximately 9·4% of all cancer deaths in 2020.[1] The disease can be triggered by histopathological changes associated with key molecular variations, such as genetic mutations, copy number alterations, and protein expressions.[2] *TP53* and *KRAS* mutations are known to drive colon cancer progression with a strong association with therapy resistance.[2] Microsatellite instability is also a key prognostic marker and is characterised by a defective DNA mismatch repair system.[3] In addition, the advent of high-quality proteomic profiles has identified functional cellular markers that cannot be reliably captured through genomic analysis.[3,4] Proteomics analysis could therefore extend the landscape of cancer genomics for differential biomarker discovery.[4] Together these multiscale molecular profiles could provide a more comprehensive view of cancer evolution to enable better patient staging, prognostication, and targeted therapy.[2]

Multilevel molecular characteristics are known to exhibit spatial differences in the tumoural micro-environment.[5] Molecularly different, spatially intertwined regions within a tumour can have individual mutational

Research in context

Evidence before this study

We searched Google Scholar and PubMed without language restrictions using the search terms "genetic mutation", "copy number alteration", "microsatellite instability", "protein expression", 'deep learning", "graph convolutional networks", "biomarkers", and "colon cancer" before doing this study. We analysed nearly 60 conference and journal articles published between Jan 1, 2018, and March 21, 2022. Previous deep-learning studies have shown their prediction capability for microsatellite instability and genetic driver mutations directly from histopathological images on colon cancer. However, these studies all analysed the image tiles separately, ignoring the importance of spatial association among tiles. In addition, existing studies seldom address cross-scale molecular profiles in colon cancer and seldom evaluate performance across cancer types and image formats. Although deep learning models have been increasingly made use of, existing publications have not assessed the spatial features of histopathological tiles in depth towards molecular profile prediction, which could have prognostic value.

Added value of this study

Tumour microenvironments have strong regional differences in image contents, so we hypothesised that the interactions between image tiles are key to understanding molecular outcomes. In this study, we present a graph neural network framework that allows the identification of multiregion spatial connection between tiles to predict cross-scale molecular profile status in colon cancer. We showed the validity of spatially connecting tumour tiles by use of the geometric coordinates from raw whole-slide images. We visualised the image tiles and measured the topological structure of tile-connected graphs. The findings expanded our understanding of histopathological characteristics with links to a large panel of cross-scale molecular profiles from genetic mutations and copy number alterations, to functional protein expressions of treatment relevance.

Implications of all the available evidence

The proposed graph neural network model is a unique way to characterise the spatial heterogeneity of the colon cancer microenvironment and has the potential to uncover widespread correlations between imaging and molecular data which can affect treatment decisions and patient prognosis and could improve management of colon cancer. The proposed graph neural network models can potentially identify multiscale molecular biomarkers for people with colon cancer, meaning that pathologists could be faster at treatment decision making, and people with colon cancer might avoid the need for molecular analysis of gene sequencing. Furthermore, the graph-based model in this study could be applied to other diseases.

outcomes, indicating the existence of intratumoural heterogeneity.[5] This spatial heterogeneity fully explains the diverse distribution of tumoural molecular subpopulations, which reflect a tumour's differential sensitivity to treatment.[6] To date, the major evidence of spatial heterogeneity comes from transcriptional and genetic profiles through the use of physically isolated biopsies on a single tumour. To discern such molecular variations, high-resolution whole-slide imaging is a unique way to assess the spatial context of the entire tumoural microenvironment (eg, cancer cells and their surrounding tissues) and tissue interactions. Identification of these histopathological patterns that are sensitive to underlying molecular mechanisms is crucial to improve our biological understanding and make more informed diagnoses. Nevertheless, quantitative whole-slide imaging that integrates regional and spatial contexts has not been investigated in depth in relation to cancer molecular signatures.

The recent development in image-genome studies has redefined the role of pathological imaging, with its capacity to characterise mutational outcomes across lung cancer,[7] liver cancer,[8] breast cancer,[6] and colon cancer,[9] among others.[10,11] This new approach is driven by our evolving ability to explore multilevel molecular profiles, such as key mutations, DNA-level copy number alterations, and functional proteomic data. Meanwhile, given the complex characteristics of tumour appearance, we need a way to automatically discover the discriminating imaging features. Deep-learning methods such as convolutional neural networks have been shown to be effective for image-based feature discovery; however, these methods are unable to directly characterise the underlying spatial information of tumoural subregions and their interactions. Considering the tumour microenvironment and its strong regional differences of image contents, we hypothesised that the differences between image tiles are key to understanding the status of molecular outcomes. Our primary objective is therefore to develop a tile-based graph analysis that could discover differential spatial characteristics from slides to help assess molecular variation, and potentially help predict clinical outcomes and targeted therapy for patients with colon cancer.

See Online for appendix

Methods

Study design

In this retrospective multicentre cohort study, we developed a spatially aware graph neural network model to predict the cross-level molecular profiles of genetic mutations, copy number alterations, and functional protein expressions from whole-slide images (appendix p 21). We designed an image-to-graph transformation that converts entire whole-slide images into a spatially connected graph representation, where the spatial connections of tumour tiles are uniquely built by use of

the geometric coordinates from the raw whole-slide images. Next, our graph neural network model architecture consists of five main modules (appendix pp 4–5), including a graph-based feature extractor, jumping knowledge structure, graph-level READOUT operation, multilayer perceptron classifier, and model ensemble strategy. We trained the proposed model on The Cancer Genome Atlas colon adenocarcinoma (TCGA-COAD); TCGA-COAD was jointly developed, beginning in 2006, by the National Cancer Institution (NCI) and the National Human Genome Research Institute to predict the molecular outcome probabilities of the corresponding whole-slide images. For each subgraph model input, the input is a group of constructed spatially connected subgraphs that is generated from each whole-slide image. The model was externally validated on The Cancer Genome Atlas rectum adenocarcinoma (TCGA-READ) and Clinical Proteomic Tumor Analysis Consortium colon adenocarcinoma (CPTAC-COAD) cohorts from NCI. Furthermore, scarcity of model understanding and results interpretation have been a concern for the widespread use of deep learning in medical research. Our graph network model employs a global sort-pooling mechanism (appendix pp 6–7) to provide possible model prediction interpretations. Full details of data preprocessing, tile selection, model structure development, and prediction interpretation are in the appendix (pp 2–6). We also did ablation studies (appendix pp 7–9) to evaluate our method design strategies (eg, the number of nodes in the graph, the distance threshold for graph edge construction, subgraph ensemble strategy, the graph convolutional layer selection, and the layer aggregation strategy) and to analyse the performance of our approach compared with baseline methods, including convolutional neural network-based methods (eg, ResNet-based model).

Multi-cohort data selection

The TCGA-COAD and the TCGA-READ datasets[12] contain 459 formalin-fixed paraffin-embedded (FFPE) stained histopathology whole-slide images of colon tumours and 165 whole-slide images of rectum tumours. The CPTAC-COAD dataset contains 161 fresh-frozen whole-slide images of colon cancer tumours.[13] TCGA-READ and CPTAC-COAD served as the external validation datasets, enabling us to evaluate model generalisability to a different cancer without using transfer learning (TCGA-READ) and a different format of whole image slides (CPTAC-COAD). There are no overlapping patients between the CPTAC-COAD and TCGA-COAD datasets. The patient characteristics of the colon and rectum datasets are shown in the appendix (p 20). We identified the associated colorectal genetic mutational profiles and microsatellite instability status from Cbioportal.[14] We also collected protein-expression profiles based on the reported clinical relevance of colon cancer and rectum cancer from The Cancer Proteome Atlas (appendix p 17).[3,4] The signature of functional protein expression can be used to identify cancer progression, metastasis, and appropriate treatments, which are not faithfully reflected by genetic alterations.[3,4] Compared to genetic changes, protein-level activities are functional and closely associated with cellular biology and drug development. We provide a detailed description of molecular profiles and the label identification in the appendix (pp 1–2).

We selected whole-slide image slides according to the following criteria: (1) the slide has no blurred areas or abnormally stained tissue areas; (2) the slide has sufficient and visible tumour regions; and (3) one slide per patient comes with available information on gene mutation, copy number alteration (eg, amplifications and deletions), microsatellite instability, and proteomics. After preprocessing, we included 306 patients with 40× magnification (0·25 microns per pixel [mmp]) in TCGA-COAD. We selected the slide in 0·25 mpp due to its higher resolution than others. The same selection criteria were applied to the validation cohorts TCGA-READ and CPTAC-COAD. We included 123 patients with whole-slide image slides and associated molecular information for TCGA-READ, and 94 for CPTAC-COAD. For microsatellite instability status classification, after preprocessing for graph construction, we included 288 slides in TCGA-COAD, 112 slides in TCGA-READ, and 94 slides in CPTAC-COAD with the available microsatellite instability records. For the proteomics analysis, we obtained high-quality proteomic profiles generated by the antibody-based technique of reverse phase protein array (RPPA) from The Cancer Proteome Atlas (TCPA) database,[3] where TCPA used a replicate-based normalisation method to combine RPPA data from different slides. CPTAC-COAD was not included in this analysis as the protein data were not available on TCPA. After whole-slide image preprocessing (eg, tile extraction and tumoural tile selection) and graph construction (appendix pp 1–7), 670 901 tiles were made use of for evaluation on colon cancer and 225 146 tiles for validation on rectum cancer. We used Python for computational analysis, including model implementation, training, and evaluation. We evaluated the performance of our model by use of area under the curve (AUC) prediction scores, their 95% CIs, and student *t* test p values.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or in the submission or writing of the report.

Results

Our model had high-level performance for predicting multiple genetic mutations in the training dataset (figure 1A and 1D; appendix pp 11–12). In particular, we found that *KRAS* mutation (AUC 80·16, 95% CI 75·83–83·93) is well predicted by our approach (appendix pp 11–12). The model also had a good prediction

For **Cbioportal** see https://www.cbioportal.org

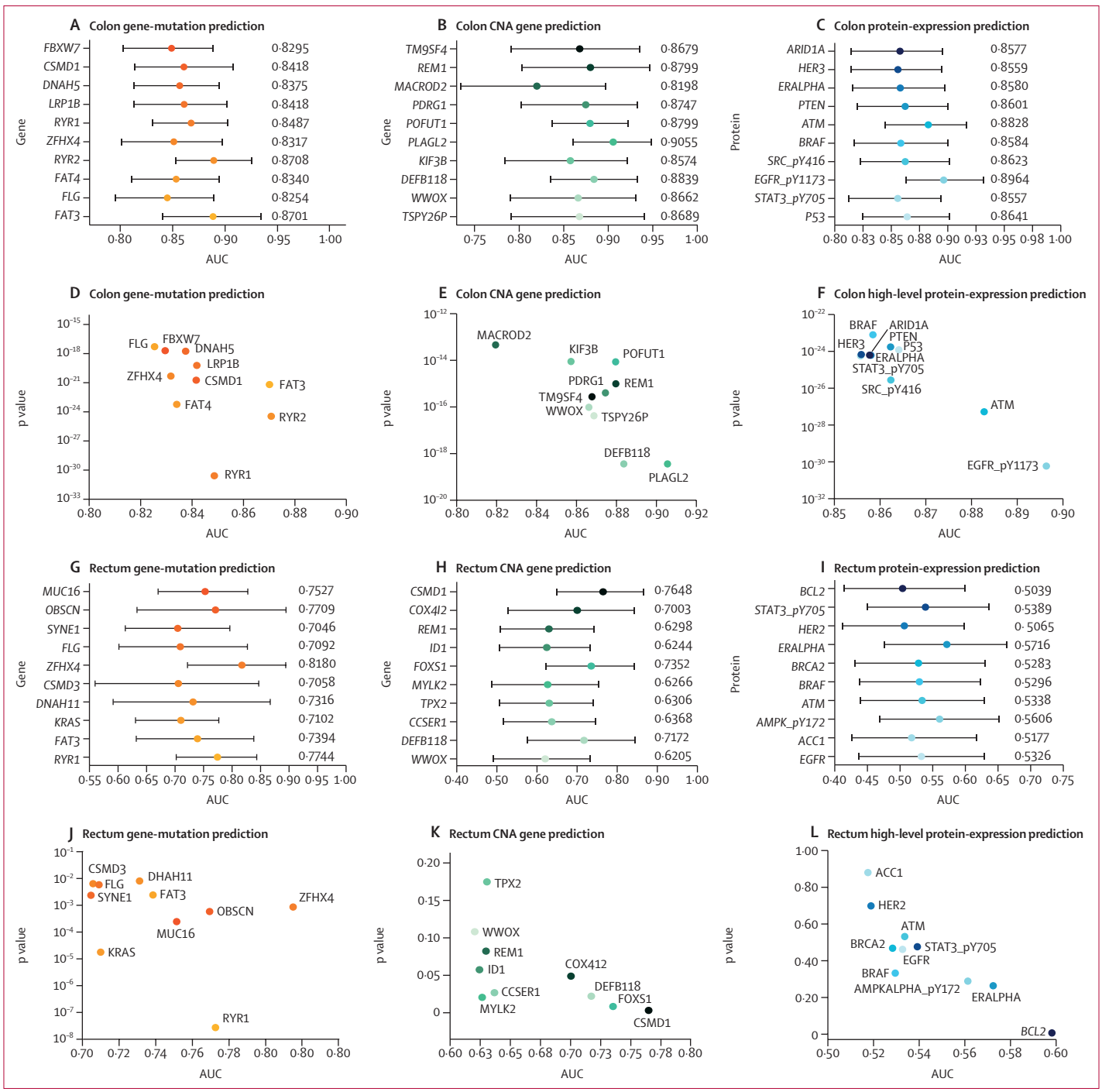


Figure 1: **Molecular profile prediction results**
The graph neural network-based model was trained to predict the molecular profile outcomes (eg, gene mutation, copy number alterations, and protein expression) on TCGA-COAD and validated on TCGA-READ. For each molecular profile, we show AUC values with student *t* test p value for the prediction scores (α=0·05). (A–C) Prediction results with 95% CI in TCGA-COAD. (D–F) Prediction results and p values in TCGA-COAD. (G–I) Prediction results with 95% CI in the TCGA-READ cohort. (J–L) Prediction results and p values in TCGA-READ. AUC=area under curve. CNA=copy number alteration. TCGA-COAD=The Cancer Genome Atlas colon adenocarcinoma. TCGA-READ=The Cancer Genome Atlas rectum adenocarcinoma.

performance for *TP53* (appendix pp 11–12) mutation (AUC 81·68, 95% CI 77·94–85·50). Our method outperformed a previous study[11,15] on a series of mutated genes (appendix pp 7–9).

Copy number alterations are a somatic change that causes the gain or loss of DNA fragments and are often associated with different cancers.[16,17] Following the same training process (figure 1B and 1D; appendix pp 13–14),

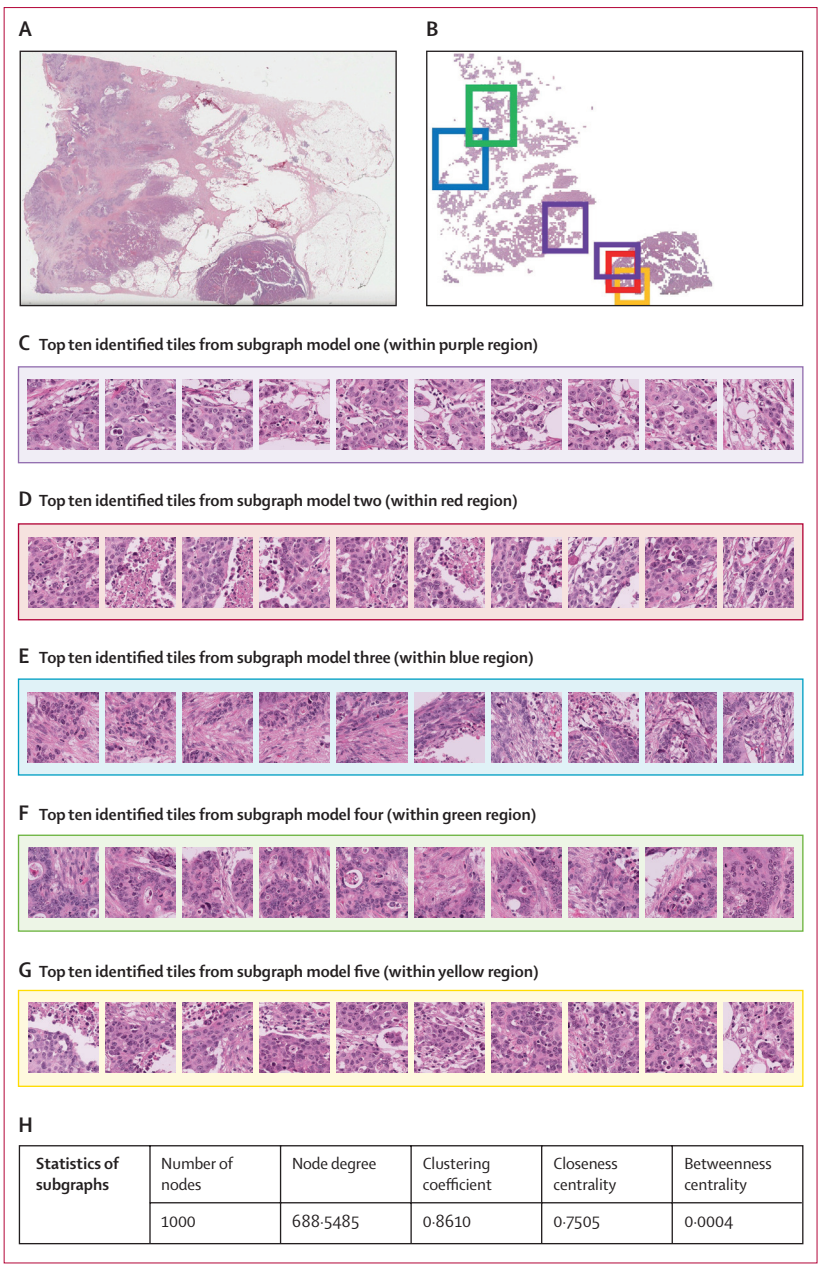
our model performed strongly (all AUCs >85·00) in predicting ten copy number alterations genes in colon cancer. For instance, both *POFUT1* (AUC 87·99, 95% CI 77·31–92·24) and *PLAG2* (AUC 90·55, 86·02–94·89) were highly predictive from our findings.

Our model also performed well for a comprehensive panel of key functional protein-expression labels in colon cancer (figure 1C and 1F; appendix pp 15–16). For instance, PTEN expression is predictable in our study (AUC 86·01, 95% CI 81·97–90·06), which represents a unique protein marker for predicting a patient's response to treatment with cetuximab.[18] The prediction result of HER3 expression (AUC 85·59, 81·39–89·48) is encouraging, since it is viewed as a determinant for poor prognosis of colon cancer.[19]

To assess the cross-cancer generalisability of the model, we externally validated it on TCGA-READ. Our model again accurately predicted multiple genetic mutations (figure 1G and 1J; appendix pp 11–12). For instance, our model could predict incidence of *KRAS* mutation (appendix pp 11–12) on rectum cancer (AUC 71·02, 95% CI 63·16–77·67), which is made use of to predict patient non-response to anti-*EGFR* target therapy (cetuximab and panitumumab).[20,21] Our model also achieved a high prediction performance for *ZFHX4* (AUC 81·80, 72·20–89·70), which is associated with poor prognosis of patients. Additionally, we found potential predictive variables of clinical relevance on the status of copy number alterations in rectum cancer (figure 1H and 1K; appendix pp 13–14).[2] Our model achieved a good prediction performance for CSMD1 (76·48, 64·78–86·71). Finally, we reported the performance of protein-expression prediction results (eg, ERALPHA 57·16, 47·54–66·46; figure 1I and 1L; appendix pp 15–16).

Our model was trained on images of FFPE slides, and so to further validate the model's potential generalisation, we tested it on CPTAC-COAD, which included fresh-frozen slides. We recognise useful findings on CPTAC-COAD to inform the model usefulness. For example, the model could predict the *DNAH5* (appendix p 18) mutation (AUC 76·16, 95% CI 67·11–83·55), which is highly associated with poor prognosis in colon cancer.[22] We also predicted the *FLG* (appendix p 18) mutation (AUC 73·45, 63·26–83·25), which is associated with loss of barrier function and deregulation of immune response.[23]

Our approach had good performance for microsatellite instability status classification in colon cancer (AUC 83·92, 95% CI 77·42–87·59). Performance is lower in rectum cancer (AUC 61·28, 53·28–67·93), probably due to the inherent image differences between cancers. We show that our microsatellite instability prediction was better (AUC 73·15, 63·21–83·13) on the CPTAC-COAD cohort despite the slide format variance. Our findings reiterate supportive evidence that predictive signals of microsatellite instability outcomes were available.[9]

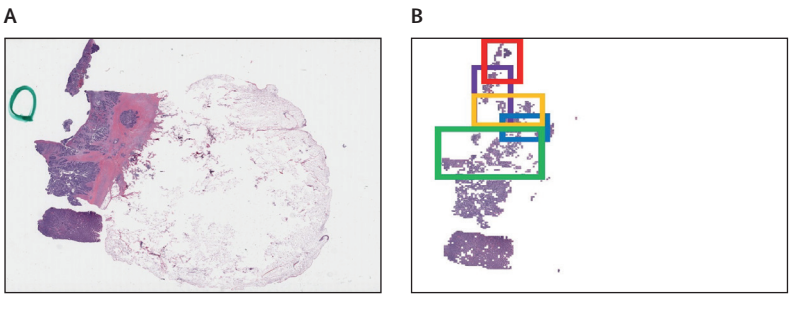


H

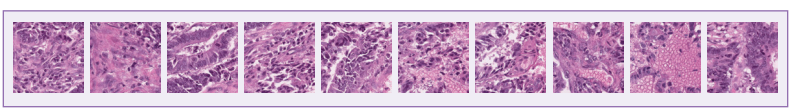
Statistics of subgraphs	Number of nodes	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	688·5485	0·8610	0·7505	0·0004

***Figure 2:* TP53 mutation prediction on TCGA-COAD**
(A) Original whole-slide imaging with *TP53* mutation outcome. (B) Highlighted regions marked by the five subgraph models within the whole-slide imaging. Different colours represent different key tile regions from subgraph models (there are two purple boxes as the titles identified by model one are spatially distributed and all regions of interest are marked). (C–G) The zoomed-in view of the identified top ten tiles from five subgraph models, which are ranked by their importance score in a decreasing order. Gross necrosis is common in tiles from model two and model three, and is rare in tiles from model one, model four, and model five. In addition, single cell necrosis is common in tiles from model one and is rare in tiles from model five. (H) Average statistical results of the graph measurements among five subgraphs. TCGA-COAD=The Cancer Genome Atlas colon adenocarcinoma.

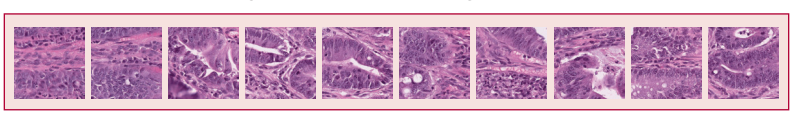
Despite the inherent differences between cancer types and image formats, we achieved a set of repeatable findings. We achieved positive gene mutation and copy number alteration gene prediction results, such as *ZFHX4* (TCGA-COAD AUC 83·17, 95% CI 78·00–87·98;



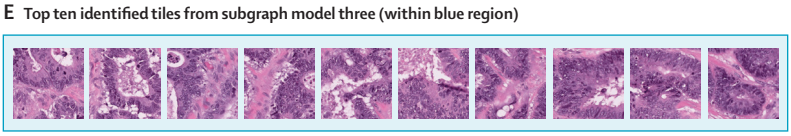
C Top ten identified tiles from subgraph model one (within purple region)



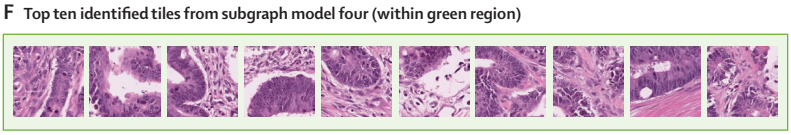
D Top ten identified tiles from subgraph model two (within red region)



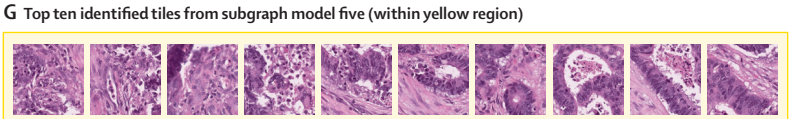
E Top ten identified tiles from subgraph model three (within blue region)



F Top ten identified tiles from subgraph model four (within green region)



G Top ten identified tiles from subgraph model five (within yellow region)



H

Statistics of subgraphs	Number of nodes	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	915·4283	0·9400	0·9177	0·0001

***Figure 3:* *PLAGL2* copy number alterations prediction on TCGA-COAD**
(A) Original whole-slide imaging with *PLAGL2* copy number alterations. (B) Highlighted regions marked by the five subgraph models within the whole-slide imaging. Different colours represent different key tile regions from subgraph models. (C–G) The zoomed-in view of the identified top ten tiles from five subgraph models, which are ranked by their importance score in a decreasing order. Tiles from models two, three, and four almost do not contain lymphocytes, and tiles from models two and three include rare apoptotic cells. Furthermore, about 40% of the tiles from model one contain single cell necrosis. (H) Average statistical results of the graph measurements among five subgraphs. TCGA-COAD=The Cancer Genome Atlas colon adenocarcinoma.

TCGA-READ AUC 81·80, 72·20–89·70) and *CSMD1* (TCGA-COAD AUC 79·86, 73·08–85·67; TCGA-READ AUC 76·48, 64·78–86·71; appendix pp 11–16, 18). Additional results include *DNAH11* (TCGA-COAD AUC 82·42, 77·16–87·75; CPTAC-COAD AUC 82·01, 74·16–88·82) and *CCSER1* (TCGA-COAD AUC 81·90, 77·16–86·54; CPTAC-COAD AUC 78·50, 67·87–87·34). Two molecular profiles could even be predicted better using our model on CPTAC-COAD than on TCGA-COAD, such as *CSMD3* (TCGA-COAD AUC 82·17, 77·82–86·57; CPTAC-COAD AUC 82·90, 73·69–90·71) and *FOXS1* (TCGA-COAD AUC 79·83, 73·18–88·14; CPTAC-COAD AUC 86·08, 79·67–91·74).

For interpretability, we display the top ten tiles with the highest contribution to the prediction based on the entire graph representation from each subgraph model (figures 2–4). We illustrated the result of *TP53* mutation by the ensemble prediction from five subgraph models, which are separately trained by tile subgraphs generated from the entire whole-slide image (figure 2).

Identified from each subgraph model, these top image tiles tend to be spatially distributed across the whole-slide image. Such a spatial characterisation is important to assess the molecular status in whole-slide images using the sort-pooling mechanism (appendix pp 7–8). The graph structure with a higher node degree and closeness centrality value (figures 3H and 4H) than the average graph statistics from across the samples (appendix p 19) is also informative as it yields accurate prediction for *PLAGL2* copy number alteration (figure 3) and PTEN protein expression (figure 4).

In our ablation analyses (appendix pp 7–9), we evaluated our method design, parameter selection, and model performance by replacing various architectures, parameters, and models in our study. We found that our subgraph ensemble strategy had the best performance across various molecular profile predictions (appendix p 34). Furthermore, our max-pooling layer aggregation in the READOUT layer could have stable performance across multiple tasks. For the graph construction strategy, the number of nodes (eg, 1000) and the distance threshold (512×85) could construct the proper graph structure for molecular profile prediction. Our study demonstrates competitive performance for individual gene mutations (appendix p 19); for example, our method consistently outperforms the previous studies on *KRAS*, *TP53*, *APC*, *TTN*, *PIK3CA*, and *FBXW7* (average AUC of 77·17). Meanwhile, our results were slightly lower than the previous study[11,15] in microsatellite instability versus stability.

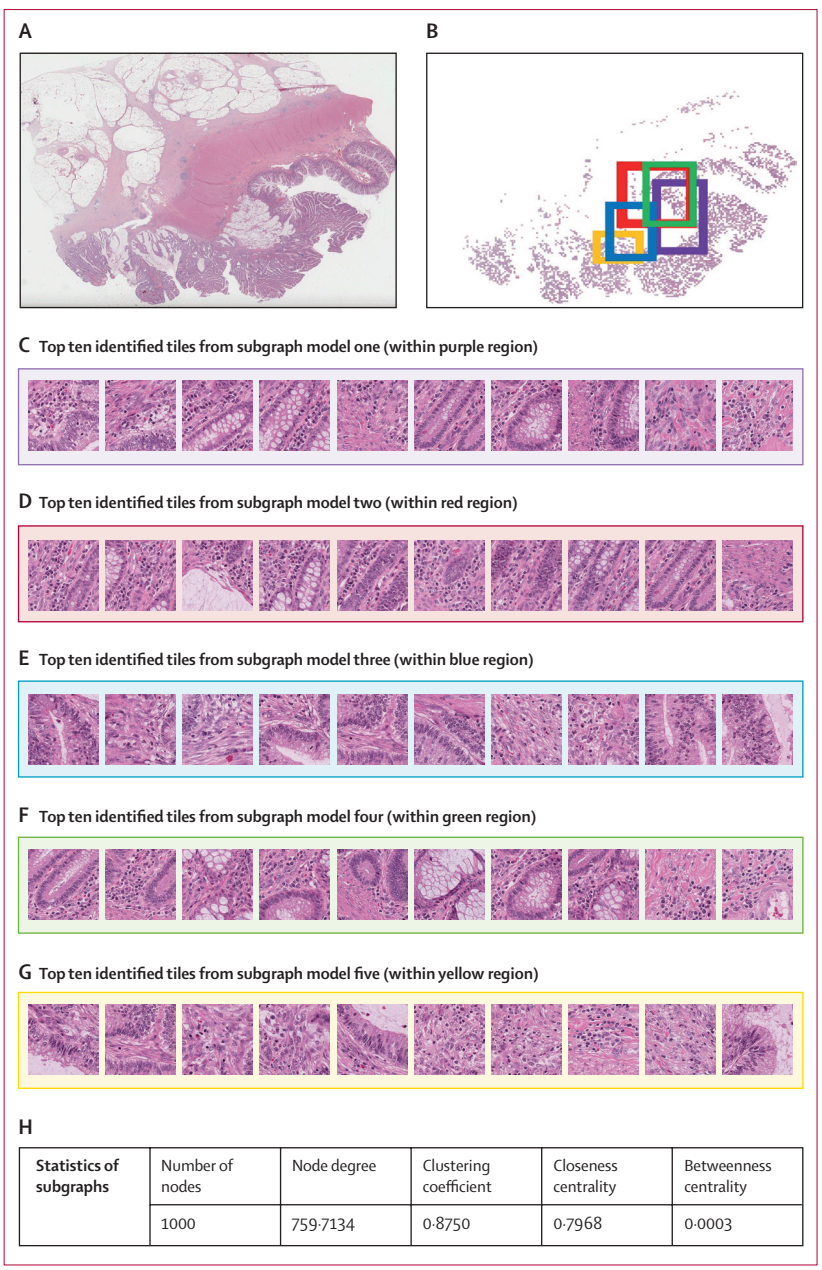
Discussion

We proposed a graph neural network approach to explore spatial information via the differences between tumoural tiles of whole-slide images. The presence of spatial and topological structures in histopathology is well documented but seldomly explored in the context of quantitative cancer imaging and machine learning.[24] Our study emphasises the use of spatial context to construct tile-connected graphs to represent histopathological slides without explicit tile annotation, which offers an efficient way to address intratumour spatial heterogeneity—crucial to our understanding of patient outcomes in colon

cancer.[5] In particular, our findings showed that a broad range of molecular-histopathological associations were predicted and may have prognostic value (eg, *KRAS* and *TP53* mutations), help assess cell progression (eg, *PLAGL2* and *POFUT1* copy number alterations), and identify targeted therapies (eg, EGFR expression) in colon cancer.

The rapid growth of whole-slide histopathology promises to uncover more meaningful genome-imaging associations via data integration.[9] Our analysis emphasises a synergistic approach in prediction and understanding of colon cancer based on molecular profiles in mutation, copy number alteration, and functional proteomics. In particular, proteomics exemplifies an emerging field to extend our understanding of genomic signatures, which permits the direct discovery of diagnostic biomarkers from a cellular cancer perspective.[25] Protein dynamics represent their own biological and cellular traits that complement roles of mRNA expression.[3,4] However, predictive analytics of proteomics profiles and their associations with other molecular signatures have not been explicitly researched in histopathology. In our study, we had good predictions on both *TP53* gene-mutation prediction (AUC 81·68, 95% CI 77·94–85·50) and P53 protein-expression prediction (AUC 86·41, 82·44–90·19). From the perspective of cancer evolution, these findings reinforce our understanding that the well known *TP53* mutation could drive the development of colon cancer with missense mutations frequently leading to the accumulation of abnormal P53 expression.[26] In addition, we identified that the expressions of NOTCH1 and copy number alterations of *POFUT1* and *PLAGL2* can be predicted because of their biological relationship.[27] Due to the challenge of cross-cancer validation, we acknowledge that the performance on the external validation set (eg, TCGA-READ and CPTAC-COAD) is reduced compared with internal validation (eg, TCGA-COAD). Our study also had a good prediction of the functional protein BRAF (AUC 85·84, 95% CI 81·68–90·03) and EGFR_pY1173 protein (AUC 89·64, 86·29–93·19), both of which are part of the EGFR-MAPK pathway, which reflects the robustness of our study as this pathway has known cancer associations. Therefore, our study makes it possible to observe cross-scale molecular activities via histopathology that were not reported in previous studies. Also, diagnosis and therapy differ considerably between colon and rectum cancers, and our results offer helpful evidence that key mutational outcomes (eg, *ZFHX4* AUC >80% and *RYR1* AUC >77% on both cancers) can be predicted, which enhances the potential clinical utility of our approach.

The image-to-graph transformation in our study allows analysis of tumoural spatial heterogeneity, as seen in histopathology. Our contributions include spatial distance definition, image-tile graph construction and labelling, and topological interpretation of spatial characteristics. Driven by the observation that spatial heterogeneity is present within and across tumoural tiles in the entire cancer microenvironment, the proposed spatial distance builds upon tiles' physical geometric coordinates to objectively capture tumoural regional differences. In addition, our tile-based graph representation enables whole-slide-level predictions,



H

Statistics of subgraphs	Number of nodes	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	759·7134	0·8750	0·7968	0·0003

Figure 4: **PTEN protein-expression prediction on TCGA-COAD**
(A) Original whole-slide imaging with PTEN protein. (B) Highlighted regions marked by the five subgraph models within the whole-slide imaging. Different colours represent different key tile regions from subgraph models. (C–G) The zoomed-in view of the identified top ten tiles from five subgraph models, which are ranked by their importance score in a decreasing order. Tiles from models 1–5 include mucinous tumour cells with background fibrosis. Clinically significant surrounding lymphocytes are included in models two, three, and four. Furthermore, single cell necrosis is visible in model five. (H) Average statistical results of the graph measurements among five subgraphs. TCGA-COAD=The Cancer Genome Atlas colon adenocarcinoma.

avoiding the uncertainty of tile label assignment for a particular molecular outcome. Such tile-based graphs do not involve extra preprocessing like nuclei or tissue segmentation, which probably brings unfavourable performance variance.[15] Assessing the full repertoire of multisized tiles is not practical, given the excessive combinations of tiles required; therefore we focused on maximising the information gathered through efficient tile samplings. To accurately show tile distribution, whole-slide tile sampling creates an unbiased space that allows for subgraph construction from the divided tumoural tiles, which enhances model generalisation and maintains a reasonable trade-off between efficiency and accuracy (appendix pp 4–5). We also provided a graph structure interpretation scheme (appendix pp 6–7) to quantitatively reveal the spatial differences between image tiles. Finally, our graph approach is purely data driven on the aggregated tumour tiles and does not rely on conventional morphological patterns that have been routinely assessed by pathologists. Consequently, the model could serve as an augmentation tool to diagnose suspicious malignancies and locate differential regions via identified tumoural tiles in histopathology.

The multigenic complexity of colon cancer is a challenge for understanding of the disease's underlying mechanisms, which makes a macroscopic approach of histopathology via powerful graph networks appealing. The strength of our graph network approach is its ability to explore the relational information among complex graph entities, which is beyond the scope of standard convolutional approaches.[28] Our analysis provides a comprehensive histopathological representation by extracting local (ie, within tile) and topological (ie, among tiles) information simultaneously, enabling a direct correlation measurement among regional tissues via importance ranking (appendix pp 1–7). The multi-parameter evaluation further reveals the stability of the proposed shallow graph neural networks (appendix p 21) across multiple prediction tasks. However, we acknowledge that there is a substantial absence of consensus guidelines on the definition and utility for tumoural image-based tiles. To address this challenge and enable detailed distribution analysis, we adopted random down-sampling with replacement to ensure that enough tiles were selected for subgraph model development.[29] Our ensemble strategy shows a simple yet effective way to merge the dynamics of tiles by aggregating prediction results between different tile-connected subgraph models.

Limitations

Although exploring the potential relationship between histopathology and molecular profiles is promising, further multisite clinical validation of our model is necessary to increase translational potential in the clinic and assist pathologists in the identification of molecular signatures in colon cancer and management of other cancers. Emerging techniques in spatial transcriptomics might provide highly defined annotations to locate fine-grained histopathological regions and further enhance deep-learning performance.[10] We recognise that the mutation imbalance (ie, the mutational rate for a particular gene across cancer types) of molecular profiles is commonly seen across cancers, making the training samples insufficient to optimise model development. For example, copy number alteration genes such as *TM9SF4*, *TPX2*, *TSPY26P*, and *WWOX* only have about 7·69% mutation ratio in the TCGA-COAD cohort, despite them having meaningful clinical relevance in colon molecular pathology.[30] We recognise that data format differences of histopathology can affect model robustness for certain mutational outcome predictions. Developing data-efficient models is therefore of interest to obtain reproducible findings in different image cohorts. Extending our graph analysis into the pan-cancer setting by assessing model consistency across cancer types would also be meaningful. Our slide inclusion criteria maintained a high quality of samples for model training; however, this stringency means that our approach might be too sensitive to slides with small artifacts. Considering the scarcity of data, we have not analysed the joint molecular activity prediction that could give knowledge about measuring complex image–genome relationships. The landscape of molecular, pathological, and predictive studies of cancer is changing rapidly, and the continued investigation of modelling imbalanced characteristics of molecular classes will be crucial to uncover additional insights into genome–pathology associations in cancer. Our findings show that exploration of the spatial characteristics of whole-slide images can well predict the cross-level molecular outcomes of patients with colon cancer.

For a **detailed introduction to the used code** see https://github.com/Cassie07/Review_Molecular_profile_prediction_GNN

For the **Genomic Data Commons portal** see https://portal.gdc.cancer.gov/

For the **Cancer Imaging Archive** see https://wiki.cancerimagingarchive.net/display/Public/Wiki

For **cBioPortal** see https://www.cbioportal.org/

For the **Cancer Proteome Atlas** see https://tcpaportal.org/tcpa/index.html

Contributors

KD and MZ designed the study and verified the data. KD contributed to data collection (eg, images and molecular profiles), data preprocessing, and data analysis. KD did the statistical analysis, which was reviewed by MZ. HW contributed to the clinical interpretation of the data and the results. All authors had access to the underlying data. SZ and DNM supervised the study. All authors interpreted the results. All authors read and approved the final version of the manuscript. All authors were responsible for the decision to submit the manuscript for publication.

Declaration of interests

MZ is an employee of Sensebrain Research, San Jose and SZ is affiliated with SenseTime Research. All other authors declare no competing interests.

Data sharing

The code used is accessible, along with a detailed introduction to the code. The whole-slide images and clinical data used in this study are publicly available through the Genomic Data Commons portal and the Cancer Imaging Archive. The multiomics data (ie, gene mutation, copy number alteration, and protein-expression data) are publicly available through cBioPortal. The functional proteomics profiles are publicly available through the Cancer Proteome Atlas and the download links of the data made use of are provided in the appendix.

Acknowledgments

This research has been partially funded by government agencies under grant numbers: ARO MURI 805491, NSF IIS-1793883, NSF CNS-1747778, NSF IIS 1763523, DOD-ARO ACC-W911NF, and NSF OIA-2040638 to DNM.

References

1 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; **68:** 394–424.

2 Armaghany T, Wilson JD, Chu Q, Mills G. Genetic alterations in colorectal cancer. *Gastrointest Cancer Res* 2012; **5:** 19–27.

3 Li J, Lu Y, Akbani R, et al. TCPA: a resource for cancer functional proteomics data. *Nat Methods* 2013; **10:** 1046–47.

4 Akbani R, Ng PKS, Werner HM, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun* 2014; **5:** 3887.

5 Ramón Y Cajal S, Sesé M, Capdevila C, et al. Clinical implications of intratumor heterogeneity: challenges and opportunities. *J Mol Med (Berl)* 2020; **98:** 161–77.

6 Naik N, Madani A, Esteva A, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nat Commun* 2020; **11:** 5727.

7 Yu K-H, Wang F, Berry GJ, et al. Classifying non-small cell lung cancer histopathology types and transcriptomic subtypes using convolutional neural networks. *bioRxiv* 2019; published online Jan 25. https://doi.org/10.1101/530360 (preprint).

8 Chen M, Zhang B, Topatana W, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol* 2020; **4:** 14.

9 Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25:** 1054–56.

10 Vickovic S, Eraslan G, Salmén F, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019; **16:** 987–90.

11 Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* 2020; **1:** 789–99.

12 Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015; **19:** A68–77.

13 Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013; **26:** 1045–57.

14 Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013; **6:** pl1.

15 Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health* 2021; **3:** e763–72.

16 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458:** 719–24.

17 Oliveira DM, Santamaria G, Laudanna C, et al. Identification of copy number alterations in colon cancer from analysis of amplicon-based next generation sequencing data. *Oncotarget* 2018; **9:** 20409–25.

18 Frattini M, Saletti P, Romagnani E, et al. PTEN loss of expression predicts cetuximab efficacy in metastatic colorectal cancer patients. *Br J Cancer* 2007; **97:** 1139–45.

19 Reschke M, Mihic-Probst D, van der Horst EH, et al. HER3 is a determinant for poor prognosis in melanoma. *Clin Cancer Res* 2008; **14:** 5188–97.

20 Soulières D, Greer W, Magliocco AM, et al. KRAS mutation testing in the treatment of metastatic colorectal cancer with anti-EGFR therapies. *Curr Oncol* 2010; **17** (suppl 1)**:** S31–40.

21 Uhlyarik A, Piurko V, Papai Z, et al. EGFR protein expression in KRAS wild-type metastatic colorectal cancer is another negative predictive factor of the cetuximab therapy. *Cancers (Basel)* 2020; **12:** 614.

22 Qing T, Zhu S, Suo C, Zhang L, Zheng Y, Shi L. Somatic mutations in ZFHX4 gene are associated with poor overall survival of Chinese esophageal squamous cell carcinoma patients. *Sci Rep* 2017; **7:** 4951.

23 Ge W, Hu H, Cai W, et al. High-risk stage III colon cancer patients identified by a novel five-gene mutational signature are characterized by upregulation of IL-23A and gut bacterial translocation of the tumor microenvironment. *Int J Cancer* 2020; **146:** 2027–35.

24 Noorbakhsh J, Farahmand S, Foroughi Pour A, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun* 2020; **11:** 6367.

25 Powers AD, Palecek SP. Protein analytical assays for diagnosing, monitoring, and choosing treatment for cancer patients. *J Healthc Eng* 2012; **3:** 503–34.

26 Williams DS, Mouradov D, Browne C, et al. Overexpression of TP53 protein is associated with the lack of adjuvant chemotherapy benefit in patients with stage III colorectal cancer. *Mod Pathol* 2020; **33:** 483–95.

27 Li D, Lin C, Li N, et al. PLAGL2 and POFUT1 are regulated by an evolutionarily conserved bidirectional promoter and are collaboratively involved in colorectal cancer by maintaining stemness. *EBioMedicine* 2019; **45:** 124–38.

28 Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst* 2021; **32:** 4–24.

29 Cook TD, Campbell DT, Shadish W. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin, 2002.

30 Żelazowski MJ, Płuciennik E, Pasz-Walczak G, Potemski P, Kordek R, Bednarek AK. WWOX expression in colorectal cancer—a real-time quantitative RT-PCR study. *Tumour Biol* 2011; **32:** 551–60.