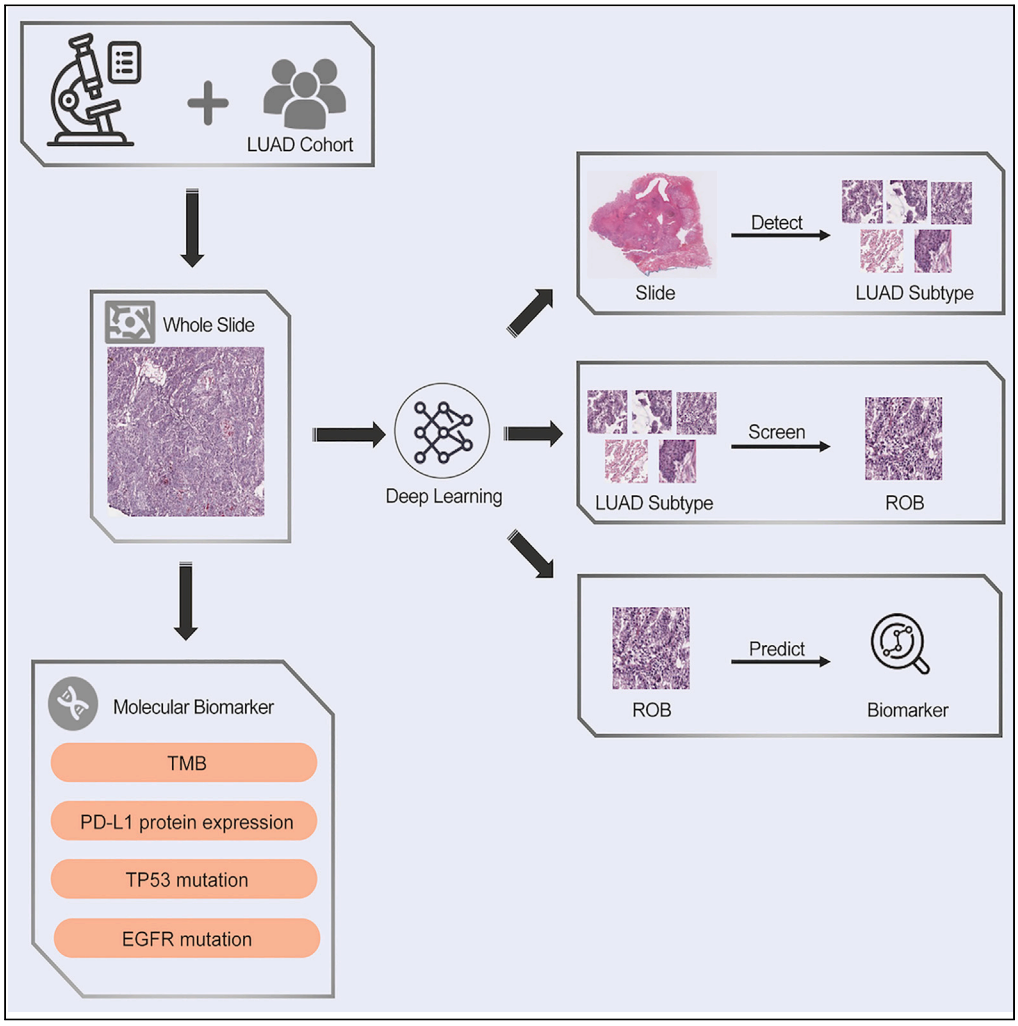
iScience

CellPress
OPEN ACCESS

Article

# Focalizing regions of biomarker relevance facilitates biomarker prediction on histopathological images




Jiefeng Gan, Hanchen Wang, Hui Yu, ..., Yaobing Chen, Guoping Wang, Tian Xia

xgwang@hust.edu.cn (X.W.)
2013tj0513@hust.edu.cn (Y.C.)
wanggp@hust.edu.cn (G.W.)
tianxia@hust.edu.cn (T.X.)

## Highlights

SRS was proposed to predict molecular biomarker status based on ROB concept

SRS can adaptively discover ROB regions targeting specific biomarkers

The generalization performance of SRS was verified on four biomarkers

Gan et al., iScience 26, 107243
October 20, 2023 
https://doi.org/10.1016/j.isci.2023.107243

iScience

CellPress
OPEN ACCESS

Article

# Focalizing regions of biomarker relevance facilitates biomarker prediction on histopathological images

Jiefeng Gan,[1,16,17] Hanchen Wang,[2,3,17] Hui Yu,[4,17] Zitong He,[5] Wenjuan Zhang,[6] Ke Ma,[1] Lianghui Zhu,[7] Yutong Bai,[5] Zongwei Zhou,[5] Alan Yullie,[5] Xiang Bai,[1,8] Mingwei Wang,[9] Dehua Yang,[9] Yanyan Chen,[10] Guoan Chen,[11] Joan Lasenby,[2] Chao Cheng,[12] Jia Wu,[13] Jianjun Zhang,[14,15] Xinggang Wang,[7,*] Yaobing Chen,[1,*] Guoping Wang,[1,*] and Tian Xia[1,16,18,*]

## SUMMARY

**Image-based AI has thrived as a potentially revolutionary tool for predicting molecular biomarker statuses, which aids in categorizing patients for appropriate medical treatments. However, many methods using hematoxylin and eosin-stained (H&E) whole-slide images (WSIs) have been found to be inefficient because of the presence of numerous uninformative or irrelevant image patches. In this study, we introduced the region of biomarker relevance (ROB) concept to identify the morphological areas most closely associated with biomarkers for accurate status prediction. We actualized this concept within a framework called saliency ROB search (SRS) to enable efficient and effective predictions. By evaluating various lung adenocarcinoma (LUAD) biomarkers, we showcased the superior performance of SRS compared to current state-of-the-art AI approaches. These findings suggest that AI tools, built on the ROB concept, can achieve enhanced molecular biomarker prediction accuracy from pathological images.**

## INTRODUCTION

Molecular biomarkers can help elucidate the diversity in disease progression and facilitate optimal therapeutic decisions, which are particularly helpful for precision treatments in cancer immunotherapy.[1–3] Meanwhile, rapidly evolving deep learning techniques[4,5] have showcased their abilities to predict biomarker statuses from histopathological whole-slide images (WSIs)[6–16] by capturing the morphological expressions caused by specific genetics and somatic mutations. Most AI-based biomarker prediction pipelines[6,8–12] operate as follows. First, they divide a whole image into a collection of patches using a fixed-size sliding window (hereafter referred to as sliding window–generated patches) and then employ a convolutional neural network (CNN) to classify each patch. The ensemble of patch predictions is subsequently aggregated using simple rules or lightweight machine learning algorithms (e.g., majority vote, logistic regression, support vector machine) to produce final predictions.

A major challenge with this whole-slide–based strategy is that it allocates equal importance to every patch, neglecting the role of varying tissue patterns. This can hinder prediction performance and may even result in failure if the image contains too many patches with unrelated tissues. For instance, a variety of morphic information, such as tumor stroma, blood vessels, and fibroblasts, has little correlation with molecular biomarker statuses. Consequently, using all the sliding window–generated patches for the classifier can introduce substantial noises and be detrimental to diagnoses during both training and inference phases. Moreover, some studies[7,17] have tried to capture tumor-containing patches as the input for the final classifier by performing multiple classifications on sliding window–generated patches. However, determining standard cancer-related labels (e.g., tumor or non-tumor) becomes challenging when various tissue patterns coexist, especially for cancers that comprise a mix of multiple histological subtypes. Furthermore, it is not ideal to develop one classifier to predict many molecular biomarkers using identical patches without adaptive selection. Distinct biological properties and causative factors accompany these biomarkers, resulting in variable and context-dependent morphological representations.

[1]Institute of Pathology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]Department of Engineering, University of Cambridge, Fitzwilliam House 32 Trumpington Street, Cambridge CB2 1QY, UK

[3]Computing + Mathematical Sciences Department, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA

[4]Wuhan Children's Hospital, Tongji Medical College, Wuhan, Hubei 430000, China

[5]Department of Computer Science, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218, USA

[6]Department of Pathology, Maternal and Child Hospital of Hubei Province, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 43000, China

[7]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei 430000, China

[8]Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 43000, China

[9]The National Center for Drug Screening, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

[10]Department of Information Management, Tongji Hospital, Huazhong

*Continued*

To address the challenges encountered by current CNN-based approaches, we propose a new morphological concept known as regions of biomarker relevance (ROB) to represent areas within a pathological image that are mostly associated with a specific biomarker. Our inventive CNN framework is built on the ROB concept, allowing for the identification of molecular biomarker statuses from H&E WSIs. Our proposed framework, termed saliency ROB Search (SRS), facilitates the direct detection of bounding boxes surrounding tumor subtypes, bypassing the need for the classification of each clipped patch. Furthermore, it identifies salient ROBs to reconstruct the patient's morphological feature representation, supplanting the original WSI for the prediction of target biomarker statuses. Utilizing WSIs from LUAD patients, we demonstrate the practicality and adaptability of SRS by examining various categories of molecular biomarkers, encompassing TMB, PD-L1 protein expression, TP53, and EGFR mutations. We are confident that this approach represents a promising avenue for further exploration of molecular biomarker status prediction using H&E WSIs.

## RESULTS

In this study, we introduce a novel deep learning framework, saliency ROB search (SRS), designed to predict molecular biomarker status based on H&E WSIs. We outline the details of SRS in the STAR Methods section. In essence, SRS comprises three cascading modules (Figure 1), the Tumor Search Module (TSM), responsible for detecting tissue of multiple histological tumor subtypes; the ROB Search Module (RSM), which filters out discriminative ROBs from predicted subtype lesions; and the Status Prediction Module (SPM) for predicting biomarker status. SRS begins with subtype lesion detection and then *in situ* searching for ROBs for every biomarker without human supervision. The adaptive ROB selection enables SRS to prioritize WSI patches most relevant to biomarkers, consistently enhancing prediction performance. To train and validate SRS, we curated and annotated four publicly available datasets, comprising 1454 H&E WSIs from 775 LUAD patients, focusing on three categories of biomarkers – TMB, PD-L1 protein expression, TP53, and EGFR mutations.

### Performance on TMB prediction

In predicting the TMB biomarker from the TCGA dataset, SRS achieved state-of-the-art (SOTA) performance, with an average precision (AP) of 0.782 (95% CI = 0.696–0.865, p = 0.045) and a better area under the curve (AUC) comparing to SOTA of 0.833 (95% CI = 0.721–0.921, p = 0.040). We validated the effectiveness of the TSM and RSM modules by varying the inputs for SPM. A detailed performance comparison is presented in Table 1. We first describe the configurations of the comparing baseline ("SRS, w/o TSM and RSM" in Table 1). In line with previous studies,[4,6,7] we utilized a sliding window of 512 × 512 pixels at a 20× objective lens magnification to partition the WSIs, preparing the input patches for the subsequent SPM module, which treats the task of predicting biomarker status as a traditional classification problem. Before deriving the outputs of SPM for patients, we used the prediction probabilities as the confidence scores to remove those ambiguous patches whose scores lie between 0.3 and 0.7. In comparison to this baseline workflow, the AUCs improved significantly by 14.2% (from 0.691 to 0.789) and 7.8% (from 0.691 to 0.745) when TSM and RSM were added correspondingly, while the rest pipeline (e.g., SPM) unchanged. We provided the ROC and PR curves of experiments w/o TSM and RSM in Figure S1. The TSM and RSM are expected to be helpful in identifying subtype lesions and eliminating regions that have little association with the target biomarker or exhibit inconsistent biomarker status compared to the source patient. Furthermore, we integrated the screened-out patches in RSM to re-train the SPM, achieving an AP of 0.538 (95% CI = 0.395–0.610, p = 0.057) and an AUC of 0.664 (95% CI = 0.552–0.730, p = 0.050), which was worse than the baseline, supporting the claims that the SRS can distinguish between meaningful and meaningless patches. The enhancement brought by TSM can be attributed to the presence of morphological signals for molecular biomarker prediction in certain tumor tissues. Meanwhile, RSM proved beneficial for prediction by discarding tissue parts (even tumor tissues) that are indistinct or misleading.

The advantages of SRS extend beyond performance to the robustness of the threshold choices for determining whether TMB status is high. Mika S. Jain et al.[4] observed that AUCs/APs experienced a noticeable degradation from 0.810/0.740 to 0.760/0.780, 0.810/0.710, and 0.700/0.410 when the threshold changed from 206 to 135.5, 223, and 293.5, respectively. These thresholds correspond to the median, tertile, and quartile of patients' TMB values. However, the issue of poor robustness has been improved due to the search strategy for salient ROBs in the SRS. As shown in Table 2, our SRS achieved AUCs of 0.785 (95% CI = 0.675–0.892, p = 0.043), 0.833 (95% CI = 0.721–0.921, p = 0.040), 0.841 (95% CI = 0.747–0.931, p = 0.042), and 0.829 (95% CI = 0.710–0.928, p = 0.040) for the same set of threshold values of 135.5, 206,

University of Science and Technology, Wuhan, Hubei 430000, China

[11]Wuhan Blood Center, Wuhan, Hubei 43000, China

[12]Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

[13]Department of Radiation Oncology, Stanford University School of Medicine, 875 Blake Wilbur Dr, Palo Alto, CA 94304, USA

[14]Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[15]Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[16]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

[17]These authors contributed equally

[18]Lead contact

*Correspondence: xgwang@hust.edu.cn (X.W.), 2013tj0513@hust.edu.cn (Y.C.), wanggp@hust.edu.cn (G.W.), tianxia@hust.edu.cn (T.X.)

https://doi.org/10.1016/j.isci.2023.107243

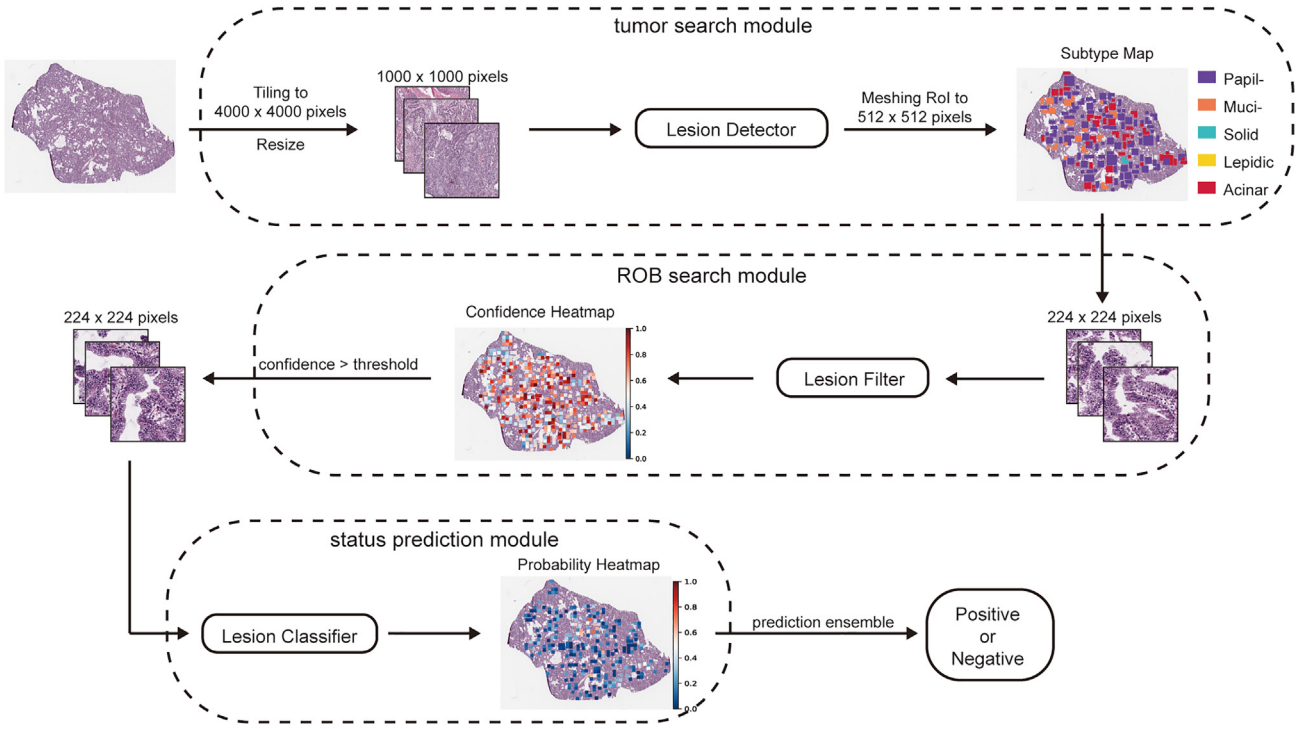



**Figure 1. An overview of saliency ROB Search (SRS) Framework**

233, and 293.5, respectively. The approximate AUC scores indicate that our SRS can handle different thresholds, even if the clinical criteria change. In addition, we attempted to apply multi-instance learning (MIL)[18–20] to address the problem of predicting biomarker status from WSIs but encountered the erroneous case where the majority of patients were predicted as having high TMB status (see further description in the discussion section).

We further evaluated the trained SRS models on the held-out test set derived from an external dataset named CPTAC and achieved an AUC of 0.732 (95% CI: 0.632–0.922, p = 0.039) for TMB status prediction. Next, we compared different CNN architectures in SPM using the TCGA dataset. The compared CNN models included ResNet50,[21] DenseNet121,[22] and Inception-V3.[23] The Light-Xception model used in SRS had fewer parameters (0.99M) and fewer computational FLOPs (388.96M), improving AUCs by 5.6% (from 0.789 to 0.833), 5.6% (from 0.789 to 0.833), and 4.1% (from 0.800 to 0.833), compared to ResNet50, DenseNet121, and Inception V3 models, respectively. The detailed performance can be found in Table S1.

We visualized the spatial distribution of histologic subtype lesions detected by SRS (Figure 2A) and the confidence heatmap for predicted subtype lesions (Figure 2B) in predicting TMB biomarker. The confidence

**Table 1. Performance comparison between different computational workflows on predicting biomarker of TMB**

| Method | Sensitivity | Specificity | AP (95%CI) | AUC (95% CI) |
|---|---|---|---|---|
| Jain, M.S.et al. | / | / | 0.740 | 0.810 |
| Xu et al. | 0.726 | 0.679 | / | 0.742 (0.682–0.794) |
| Sadhwani, A. et al. | 0.712 | 0.717 | / | 0.770 (0.640–0.880) |
| w/o TSM and RSM | 0.667 | 0.750 | 0.422 (0.239–0.668) | 0.691 (0.502–0.851) |
| w/o TSM | 0.727 | 0.651 | 0.442 (0.300–0.723) | 0.745 (0.733–0.833) |
| w/o RSM | 0.720 | 0.735 | 0.733 (0.604–0.760) | 0.789 (0.661–0.899) |
| SRS | 0.760 | 0.816 | 0.782 (0.696–0.865) | 0.833 (0.721–0.921) |

Here we directly reported the results published in the related work.

**Table 2. Performance comparison of SRS between various TMB threshold values**

| Threshold | Sensitivity | Specificity | AP (95%CI) | AUC (95% CI) |
|---|---|---|---|---|
| 135.5 | 0.714 | 0.692 | 0.802 (0.682–0.883) | 0.785 (0.675–0.892) |
| 206 | 0.760 | 0.816 | 0.782 (0.696–0.865) | 0.833 (0.721–0.921) |
| 223 | 0.750 | 0.740 | 0.771 (0.690–0.859) | 0.841 (0.747–0.931) |
| 293.5 | 0.727 | 0.712 | 0.771 (0.670–0.868) | 0.829 (0.710–0.928) |

score indicated the model's uncertainty about its prediction. Patches with high confidence scores demonstrated a stronger correlation with the TMB biomarker. In addition, the problem of lacking accurate labels for patches is a common challenge in WSI-based methods. During training the RSM, we assigned the source patient's status as the supervisory signal for their associated subtype lesions as the standard WSI diagnosis method. It was not accurate enough, and whether the classification problem could be solved relied heavily on the accuracy of such pseudo-label assignment. For instance, a patient with low TMB status may have tissue patches exhibiting high TMB status on histopathological images and vice versa. That is a possible reason why the standard MIL method[24] and CLAM[25] cannot perform well in identifying patients with low TMB status. Actually, the confidence branch in RSM will reduce the adverse effect brought by such mismatched pairs. With the RSM, we can improve the model's tolerance for incorrect supervision

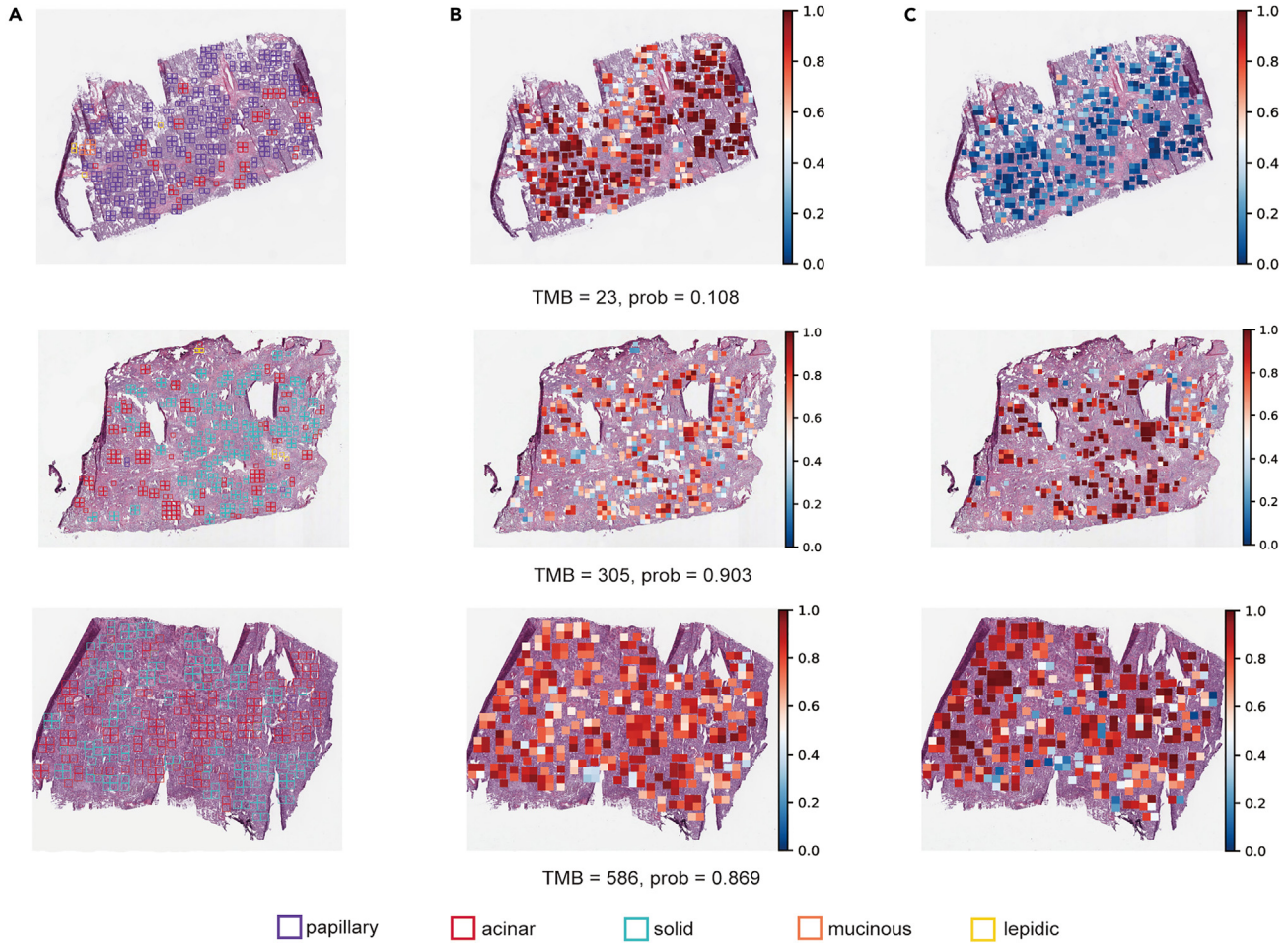



**Figure 2. Visualization of spatial heterogeneity for subtype lesions, confidence heatmap, and probability heatmap on TMB biomarker prediction**

The sampled three patients have TMB of 23, 305, and 586 (206 served as the threshold value for training).

(A) spatial distribution of subtype lesion.

(B) confidence heatmap indicating the association of lesions to TMB status.

(C) probability heatmap indicating the probability of behaving TMB-High status.

**Table 3. Performance comparison between the naive workflow and SRS on predicting PD-L1 protein expression, TP53, and EGFR mutations**

| Biomarker | | Sensitivity | Specificity | AP (95%CI) | AUC (95% CI) |
|---|---|---|---|---|---|
| PD-L1 | w/o TSM and RSM | 0.765 | 0.712 | 0.691 (0.561–0.720) | 0.803 (0.655–0.925) |
| | SRS | 0.824 | 0.772 | 0.749 (0.628–0.811) | 0.878 (0.756–0.965) |
| TP53 | w/o TSM and RSM | 0.545 | 0.558 | 0.615 (0.457–0.705) | 0.642 (0.508–0.757) |
| | SRS | 0.719 | 0.714 | 0.779 (0.689–0.857) | 0.793 (0.686–0.884) |
| EGFR | w/o TSM and RSM | 0.818 | 0.677 | 0.471 (0.225–0.549) | 0.775 (0.592–0.917) |
| | SRS | 0.778 | 0.923 | 0.756 (0.666–0.850) | 0.918 (0.782–0.995) |

by minimizing the confidence score rather than the distribution difference between the probability and pseudo-label by multiplying the confidence score with the common cross-entropy loss as a new optimization objective.

By means of selecting the subtype lesions whose confidence scores exceed a threshold, the RSM enabled recognition of the lesions contributing to TMB prediction and exhibiting TMB status consistent with the respective patient. Our SRS is capable of automatically selecting the salient ROBs, eliminating the uninformative and mismatched patches as much as possible, leading to consistent improvements compared to the current deep learning counterparts.[6,7,17] During the training of the CNN classifier in SPM, we continued to set the patient's TMB status as supervision for the lesions chosen through TSM and RSM modules. Furthermore, we provided a heatmap that indicated the probability of behaving high TMB status for each salient ROB (Figure 2C). This enabled the salient ROBs to align more closely with the predictive outcomes.

The superior performance achieved by the SRS stems from focusing the salient subtype lesions that are highly correlated with the TMB biomarker status. In addition, we conducted an analysis of the relationship between TMB status and the predominant subtype for WSIs, as well as the relationship between the two predicted attributes of TMB status and subtype category for salient ROBs. The distribution statistics were summarized in Figure S2, where the "WSIs" group described the number of WSIs exhibiting high TMB status in the subset of WSIs with a specific predominant adenocarcinoma, and the "ROBs" group 2indicated the number of ROBs predicted as high TMB status and certain subtype adenocarcinoma simultaneously. These two statistics revealed that the solid subtype has a more substantial association with high TMB status. Correspondingly, the low TMB status is concentrated in the papillary and mucinous subtypes, which aligns with previous research findings.[26]

## Performance on other biomarkers

The SRS had the adaptive ability to predict other molecular biomarkers' status with user-specified positive cutoffs. We experimentally verified such capability with a new set cutoff value of 152 on PD-L1 protein expression, TP53, and EGFR mutations, achieving AUCs of 0.878 (95% CI = 0.756–0.965, p = 0.042), 0.793 (95% CI = 0.686–0.884, p = 0.040) and 0.918 (95% CI = 0.782–0.995, p = 0.044) on the TCGA test split, respectively. We presented the model performance and dataset statistics in Tables 3 and 4, respectively. Regarding the TMB biomarker, we examined the statistical relationship between subtype and biomarker status. High PD-L1 protein expression and TP53 mutation occurred more frequently in the solid subtype, while EGFR mutations were more common in the acinar subtype. We presented a more comprehensive analysis in Figure S2.

To demonstrate the importance of the ROB concept, we present the ROB diversity of TMB, PD-L1 protein expression, TP53, and EGFR mutations in Figure 3. It is evident that each biomarker is associated with its own unique set of ROBs. In addition, when we simply fed an entire WSI as model input, the prediction performance dropped significantly, as similarly observed in TMB prediction (Figure S3). This may reflect the fundamental principle that different biomarkers have distinct biological causes at the genetic level and, consequently, display unique morphological/phenotypic representations on a WSI.

Furthermore, we compared the selected salient ROBs with response regions based on immunohistochemistry (IHC) for estimating the biomarker of PD-L1 protein expression biomarker on slides from the Wuhan Tongji Hospital. We discovered a high similarity between ROBs with a high probability of being positive (red square) and response regions delineated with red curves in Figure 4.

**Table 4. Demographical and clinical information on LUAD dataset from the TCGA, NLST, WHTJ, and CPTAC cohorts**

| Dataset | | Female | Male | Age range (median) | Slide amount | Patient amount |
|---|---|---|---|---|---|---|
| TCGA | Trainval | 149 | 143 | 33-88 (66) | 439 | 292 |
| | Test | 41 | 33 | 41-87 (65) | 111 | 74 |
| NLST | Trainval | / | / | / | 209 | 118 |
| | Test | | | | 47 | 29 |
| WHTJ | Trainval | / | / | / | 123 | 123 |
| | Test | | | | 34 | 34 |
| CPTAC | Test | 37 | 68 | / | 491 | 105 |

## DISCUSSION

Current deep learning methods for predicting molecular biomarker status from histopathology mainly were based on inputs of whole slide/whole tumor (WS/WT). However, human cancers displayed significant intra-tumor heterogeneity in morphological and phenotypic features, which could be detrimental to the WS/WT-based strategy. There were methods[17,27] proposed to recognize the representative tissue using affinity propagation clustering,[28] an unsupervised method, from tumor-containing patches. However, applying affinity propagation clustering to search salient ROBs could be inefficient because it was intricacy to determine which clusters were helpful in prediction based on some heuristic rules or experience.

We have tried to apply the standard MIL with maximum pooling to perform TMB prediction using sliding window–generated patches. Here, maximum pooling meant sampling patches from each WSI with the largest prediction probability to optimize the model at each iteration. Meanwhile, during inference, we chose the patches with the largest probability to represent the prediction of the source patient, but they were all determined to be high with a sensitivity of 1.0 and specificity of 0. In addition, CLAM[25] is an extension of the MIL framework, which incorporates instance-level clustering and attention-based pooling for accurate classification. We conducted an additional experiment to evaluate the performance of CLAM[25] in classifying TMB biomarkers using only tumor (subtype) patches, achieving a sensitivity of 0.943 and specificity of 0.133, better than the standard MIL but far behind our SRS. The two unsuccessful attempts provided support for our conjecture that there were always tissue patches that exhibited high TMB despite the source patient being under low TMB status. This discovery contradicted the fundamental bag-instance assumption in MIL, which states that negative bags should not contain any positive instances and positive bags should contain at least one positive instance. The presence of such mismatches made it challenging to optimize for predicting TMB status, similar to optimizing MIL for natural objects.

On such a basis, we proposed a framework titled SRS to capture morphological patterns that were the most correlated with the target biomarker, attempting to build an association between genotype and phenotype. This improved model performance and enabled the reveal of biological insights hidden in histopathological images. Our work demonstrated that LUAD patients' biomarker status, such as TMB, PD-L1 protein expression, and TP53 mutation, etc., can be predicted with digitalized H&E frozen WSIs, and showed an edge over conventional methods taking the entire WSI as input. Our SRS design can be applied to estimate the status of a single gene mutation and the mixed effect of multiple gene mutations. Furthermore, we

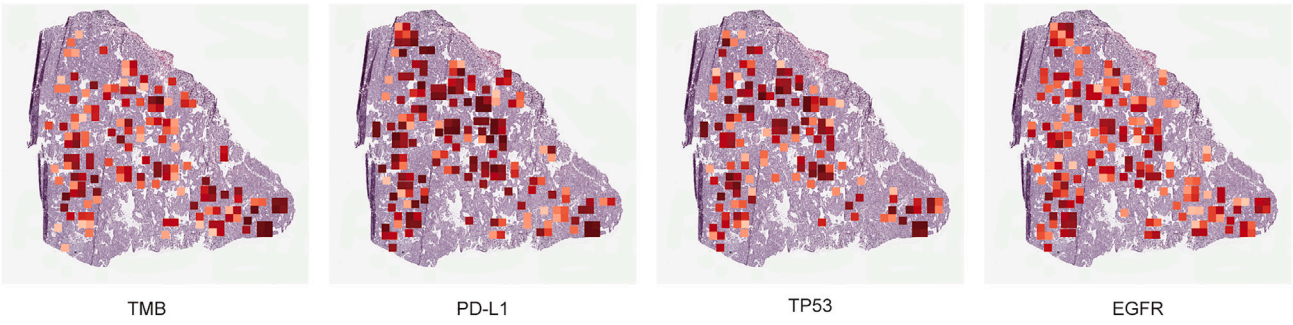



**Figure 3. The distribution of salient ROBs contributing to the prediction of biomarkers of TMB, PD-L1 protein expression, TP53, and EGFR mutations**

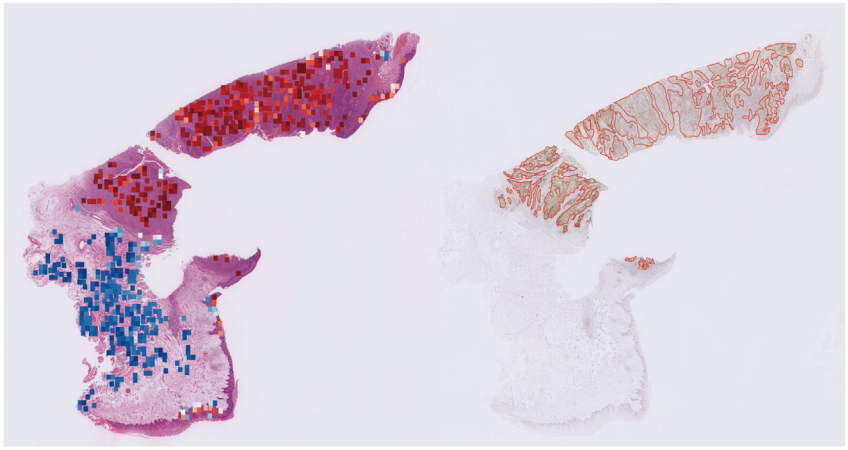


**Figure 4. The comparison between ROBs in the H&E image and response region in the IHC image (red curve) for evaluating PD-L1 protein expression**

On the left image, the red color indicates patches with a high probability of positive PD-L1 status, while the blue color indicates a low probability.

verified the robustness of the threshold choice, which showed that SRS could be workable for various biomarkers with diverse binarization settings.

### Limitations of the study

Despite the promises, there are certain limitations of SRS. One major limitation is that molecular biomarkers with continuous values (e.g., TMB) were analyzed with coarse categories instead of fine numerical values. A more quantitative analysis beyond binary classification requested more patients with continuous values varying the diverse ranges for training. Moreover, the lack of a carefully curated and comprehensive dataset poses the difficulty in developing an efficient algorithm to produce a foundational feature encoder based on high-resolution H&E WSIs, where the extracted features should be directly utilized for training biomarker value regressors. Another limitation pertains to the use of the classic Cascade R-CNN for detecting tumor subtype, which cannot represent the most advanced object detection models currently available. Nonetheless, we intend to address this limitation in further work by incorporating the latest transformer-based methods to update our detector. Despite these, our proposed SRS framework validated the venue of selecting saliency ROBs for biomarker prediction and brought substantial improvements to the current deep-learning solutions. Our study highlights the potential for further development of the SRS framework and its application in clinical practice.

## STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Image datasets
  - Dataset split
  - Image annotation
  - Biomarker processing
- METHOD DETAILS
  - Tumor search module
  - ROB search module
  - Status prediction module
  - Prediction ensemble

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107243.

## ACKNOWLEDGMENTS

The authors would like to thank the two anonymous reviewers for their valuable comments and suggestions. This work was supported by China Hainan Provincial Major Science and Technology Project ZDKJ2021028 (D.Y.), The University of Texas MD Anderson Lung Moon Shot Program, The University of Texas MD Anderson Cancer Center Core Grant P30 CA01667, the National Institutes of Health (NIH) grant R00CA218667, R01CA234629, the AACR-Johnson & Johnson Lung Cancer Innovation Science Grant (18-90-52-ZHAN), the Rexanna's Foundation for Fighting Lung Cancer, Sabin Family Fund, Rydin Family Research Fund.

## AUTHOR CONTRIBUTIONS

J.G., H.W., X.W., Y.C., G.W., and T.X. conceived and designed the study. J.G., H.W., Y.C., and T.X. did the literature search. Y.C., W.Z., M.W., D.Y., Y.C., G.C., and G.W. contributed to lung adenocarcinoma subtype annotation and immunohistochemistry (IHC) on PD-L1 biomarker. J.G., Y.C., and H.W. preprocessed the data. J.G. and H.W. did the deep learning model development and performance evaluation. J.G. analyzed and interpreted the data, and drafted the manuscript. H.W., Z.H., K.M., H.Z., Y.B., Z.Z., A.Y., B.X., J.L., X.G., C.C., J.W., J.Z., and T.X. critically revised the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

## DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: February 14, 2023
Revised: May 11, 2023
Accepted: June 26, 2023
Published: June 29, 2023

## REFERENCES

1. Darvin, P., Toor, S.M., Sasidharan Nair, V., and Elkord, E. (2018). Immune checkpoint inhibitors: recent progress and potential biomarkers. Exp. Mol. Med. *50*, 1–11. https://doi.org/10.1038/s12276-018-0191-1.

2. Chan, T.A., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S.A., Stenzinger, A., and Peters, S. (2019). Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. Ann. Oncol. *30*, 44–56. https://doi.org/10.1093/annonc/mdy495.

3. Hargadon, K.M., Johnson, C.E., and Williams, C.J. (2018). Immune checkpoint blockade therapy for cancer: an overview of FDA-approved immune checkpoint inhibitors. Int. Immunopharmacol. *62*, 29–39. https://doi.org/10.1016/j.intimp.2018.06.001.

4. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

5. Bai, X., Wang, H., Ma, L., Xu, Y., Gan, J., Fan, Z., Yang, F., Ma, K., Yang, J., Bai, S., et al. (2021). Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. Nat. Mach. Intell. *3*, 1081–1089. https://doi.org/10.1038/s42256-021-00421-z.

6. Jain, M.S., and Massoud, T.F. (2020). Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nat. Mach. Intell. *2*, 356–362. https://doi.org/10.1038/s42256-020-0190-5.

7. Sadhwani, A., Chang, H.-W., Behrooz, A., Brown, T., Auvigne-Flament, I., Patel, H., Findlater, R., Velez, V., Tan, F., Tekiela, K., et al. (2021). Comparative analysis of machine learning approaches to classify tumor mutation burden in lung adenocarcinoma using histopathology images. Sci. Rep. *11*, 16605. https://doi.org/10.1038/s41598-021-95747-4.

8. Wang, L., Jiao, Y., Qiao, Y., Zeng, N., and Yu, R. (2020). A novel approach combined transfer learning and deep learning to predict TMB from histology image. Pattern Recognit. Lett. *135*, 244–248. https://doi.org/10.1016/j.patrec.2020.04.008.

9. Murchan, P., Ó'Brien, C., O'Connell, S., McNevin, C.S., Baird, A.-M., Sheils, O., Ó Broin, P., and Finn, S.P. (2021). Deep learning of histopathological features for the prediction of tumour molecular genetics. Diagnostics *11*, 1406. https://doi.org/10.3390/diagnostics11081406.

10. Sha, L., Osinski, B.L., Ho, I.Y., Tan, T.L., Willis, C., Weiss, H., Beaubier, N., Mahon, B.M., Taxter, T.J., and Yip, S.S.F. (2019). Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. J. Pathol. Inform. *10*, 24. https://doi.org/10.4103/jpi.jpi_24_19.

11. Noorbakhsh, J., Farahmand, S., Foroughi pour, A., Namburi, S., Caruana, D., Rimm, D., Soltanieh-ha, M., Zarringhalam, K., and Chuang, J.H. (2020). Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nat. Commun. *11*, 6367. https://doi.org/10.1038/s41467-020-20030-5.

12. Jang, H.-J., Lee, A., Kang, J., Song, I.H., and Lee, S.H. (2021). Prediction of genetic alterations from gastric cancer histopathology images using a fully automated deep learning approach. World J. Gastroenterol. *27*, 7687–7704. https://doi.org/10.3748/wjg.v27.i44.7687.

13. Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., and Rajpoot, N.M. (2021). 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 *3*, e763–e772. https://doi.org/10.1016/S2589-7500(21)00180-1.

14. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567. https://doi.org/10.1038/s41591-018-0177-5.

15. Chen, Y., Yang, H., Cheng, Z., Chen, L., Peng, S., Wang, J., Yang, M., Lin, C., Chen, Y., Wang, Y., et al. (2022). A whole-slide image (WSI)-based immunohistochemical feature prediction system improves the subtyping of lung cancer. Lung Cancer *165*, 18–27. https://doi.org/10.1016/j.lungcan.2022.01.005.

16. Patil, P.D., Hobbs, B., and Pennell, N.A. (2019). The promise and challenges of deep learning models for automated histopathologic classification and mutation prediction in lung cancer. J. Thorac. Dis. *11*, 369–372. https://doi.org/10.21037/jtd.2018.12.55.

17. Xu, H., Park, S., Clemenceau, J.R., Choi, J., Radakovich, N., Lee, S.H., and Hwang, T.H. (2019). Spatial heterogeneity and organization of tumor mutation burden and immune infiltrates within tumors based on whole slide images correlated with patient survival in bladder cancer. Bioinformatics. https://doi.org/10.1101/554527.

18. Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., and Huang, J. (2018). Deep multi-instance learning with dynamic pooling. In Proceedings of The 10th Asian Conference on Machine Learning (PMLR), *95*, pp. 662–677.

19. Li, B., Li, Y., and Eliceiri, K.W. (2021). Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 14313–14323. https://doi.org/10.1109/CVPR46437.2021.01409.

20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and Zhang, Y. (2021). TransMIL: transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. *34*, 2136–2147.

21. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

22. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.

24. Wang, X., Yan, Y., Tang, P., Bai, X., and Liu, W. (2018). Revisiting multiple instance neural networks. Pattern Recogn. *74*, 15–24. https://doi.org/10.1016/j.patcog.2017.08.026.

25. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. *5*, 555–570. https://doi.org/10.1038/s41551-020-00682-w.

26. Hellmann, M.D., Ciuleanu, T.-E., Pluzanski, A., Lee, J.S., Otterson, G.A., Audigier-Valette, C., Minenza, E., Linardou, H., Burgers, S., Salman, P., et al. (2018). Nivolumab plus Ipilimumab in lung cancer with a high tumor mutational burden. N. Engl. J. Med. *378*, 2093–2104. https://doi.org/10.1056/NEJMoa1801946.

27. Hu, J., Cui, C., Yang, W., Huang, L., Yu, R., Liu, S., and Kong, Y. (2021). Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images. Transl. Oncol. *14*, 100921. https://doi.org/10.1016/j.tranon.2020.100921.

28. Wang, K., Zhang, J., Li, D., Zhang, X., and Guo, T. (2008). Adaptive affinity propagation clustering. Preprint at arXiv. https://doi.org/10.48550/arXiv.0805.1096.

29. Travis, W.D., Brambilla, E., Nicholson, A.G., Yatabe, Y., Austin, J.H.M., Beasley, M.B., Chirieac, L.R., Dacic, S., Duhig, E., Flieder, D.B., et al. (2015). The 2015 World Health Organization classification of lung tumors. J. Thorac. Oncol. *10*, 1243–1260. https://doi.org/10.1097/JTO.0000000000000630.

30. Fumet, J.-D., Richard, C., Ledys, F., Klopfenstein, Q., Joubert, P., Routy, B., Truntzer, C., Gagné, A., Hamel, M.-A., Guimaraes, C.F., et al. (2018). Prognostic and predictive role of CD8 and PD-L1 determination in lung tumor tissue of patients under anti-PD-1 therapy. Br. J. Cancer *119*, 950–960. https://doi.org/10.1038/s41416-018-0220-9.

31. Molica, M., Mazzone, C., Niscola, P., and de Fabritiis, P. (2020). TP53 mutations in acute myeloid leukemia: still a daunting challenge? Front. Oncol. *10*, 610820. https://doi.org/10.3389/fonc.2020.610820.

32. Marcus, L., Fashoyin-Aje, L.A., Donoghue, M., Yuan, M., Rodriguez, L., Gallagher, P.S., Philip, R., Ghosh, S., Theoret, M.R., Beaver, J.A., et al. (2021). FDA approval summary: pembrolizumab for the treatment of tumor mutational burden–high solid tumors. Clin. Cancer Res. *27*, 4685–4689. https://doi.org/10.1158/1078-0432.CCR-21-0327.

33. Budczies, J., Klauschen, F., Sinn, B.V., Győrffy, B., Schmitt, W.D., Darb-Esfahani, S., and Denkert, C. (2012). Cutoff finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. PLoS One *7*, e51862. https://doi.org/10.1371/journal.pone.0051862.

34. Cai, Z., and Vasconcelos, N. (2018). Cascade R-CNN: delving into high quality object detection. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 6154–6162. https://doi.org/10.1109/CVPR.2018.00644.

35. Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. *39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

36. Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2021). Res2Net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. *43*, 652–662. https://doi.org/10.1109/TPAMI.2019.2938758.

37. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 936–944. https://doi.org/10.1109/CVPR.2017.106.

38. Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. In In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 6517–6525. https://doi.org/10.1109/CVPR.2017.690.

39. Girshick, R. (2015). Fast R-CNN. In 2015 IEEE International Conference on Computer Vision (ICCV) (IEEE), pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169.

40. DeVries, T., and Taylor, G.W. (2018). Learning confidence for out-of-distribution detection in neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.04865.

41. Chollet, F. (2017). Xception: deep learning with depthwise Separable convolutions. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195.

42. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al. (2019). MMDetection: open MMLab detection toolbox and benchmark. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.07155.

43. DeVries, T., and Taylor, G.W. (2017). Improved regularization of convolutional neural networks with cutout. Preprint at arXiv. https://doi.org/10.48550/arXiv.1708.04552.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| The Cancer Genome Atlas (TCGA) | National Cancer Institute | https://portal.gdc.cancer.gov/ |
| The National Lung Screening Trial (NLST) | National Cancer Institute | https://cdas.cancer.gov/datasets/nlst/ |
| The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) | National Cancer Institute | https://www.cancerimagingarchive.net/ |
| Software and algorithms | | |
| OpenSlide | OpenSlide team | https://openslide.org/ |
| OpenCV | OpenCV team | https://opencv.org/ |
| Pytorch | Meta AI | https://pytorch.org/ |
| MMDetection | OpenMMLab | https://github.com/open-mmlab/mmdetection |
| SRS | This paper | https://github.com/ganjf/biomarkerPrediction |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to the lead contact, Tian Xia (tianxia@hust.edu.cn).

### Materials availability

This study did not generate any new unique reagents.

### Data and code availability

- This paper used both existing LUAD WSI from TCGA, CPTAC and NLST datasets and newly acquired LUAD WSI from WHTJ dataset to train and validate the computional framework. The accession numbers for theses publicy available datasets of TCGA, CPTAC and NLST are listed in the key resources table. The newly acquired WHTJ dataset will be shared by the lead contact upon request.
- The original code will be available at https://github.com/ganjf/biomarkerPrediction.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

### Image datasets

In this study, we have curated 1454 digitized H&E frozen WSIs from 775 LUAD patients that were collected from four resources: the Cancer Genome Atlas (TCGA), the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), the National Lung Screening Trial (NLST), and the Wuhan Tongji Hospital (WHTJ). We summarized the dataset statistics in Table 4, including the available gender and age information, with '/' denoting the inaccessible clinical information for the NLST, WHTJ, and CPTAC datasets. The TCGA dataset consists of 550 WSIs from 366 patients, whose resolutions are either 0.25 μm/pixel (40×) or 0.5 μm/pixel (20×). The CPTAC dataset consists of 491 WSIs from 105 patients, and the slides were scanned at a resolution of 0.5 μm/pixel (20×). The NLST dataset includes 256 WSIs from 147 patients scanned at resolutions of 0.25 μm/pixel (40×) or 0.5 μm/pixel (20×). The WHTJ dataset consists of 157 WSIs from 157 patients, and the slides were scanned at a resolution of 0.25 μm/pixel (40×).

### Dataset split

For TCGA, NLST, and WHTJ datasets, we utilized 80% of cohort patients for training and validation (train-val) and the rest 20% for testing, where the CPTAC dataset served as a hold-out external testing set. Due to

the lack of clinical genetic information, NLST and WHTJ datasets do not serve for experiments of biomarker prediction. Still, they are only used to validate the performance of LUAD subtype detection.

### Image annotation

Our study comprised 1454 WSIs from 775 LUAD patients in four sources. The LUAD is recommended to be classified as multiple histologic subtypes by the 2015 World Health Organization (WHO) Classification of Lung Tumors.[29] Six qualified pathologists from Wuhan Tongji Hospital, with an average of 9 years experience, reviewed and annotated all WSIs from TCGA, NLST, and WHTJ datasets, using our own developed software for WSIs labeling to outline rectangular bounding boxes of histologic adenocarcinoma subtypes. Each annotation of adenocarcinoma subtypes has been confirmed by other colleagues from the expert panel to establish a consensus.

### Biomarker processing

In this work, we validated our model on three different types of biomarkers, including ensemble somatic mutations (TMB), gene-level somatic mutations (TP53 and EGFR), and protein expression (PD-L1) on the TCGA dataset. The CD274 gene was measured to evaluate PD-L1 protein expression. The actual biomarker status was determined by the whole-exome sequencing (WES) for TMB, by RNA sequencing for CD274,[30] and by next-generation sequencing (NGS)[31] for TP53 and EGFR mutations, respectively. We plotted the distribution of TCGA patients' TMB value and PD-L1 protein expression in Figure S4. The TMB and PD-L1 protein expressions are continuous variables, ranging from zero to thousands; therefore, the status was grouped into high/positive and low/negative depending on the preset threshold value. Coherent with previous studies,[6,26] we set the threshold value as 206 mutations for TMB, which approximately equated to 10 mutations per megabase (muts/Mb) when analyzing with the FoundationOne CDx assay, as suggested in the FDA approval.[32] With such standards, 32.24% and 36.19% of patients in the TCGA and CPTAC datasets are identified with TMB-High status. The optimal cutoff value for the PD-L1 protein expression was chosen using Cutoff Finder,[30,33] a standard optimization software for biomarker cutoff determination. In the TCGA dataset, 20.22% of patients were identified as positive in PD-L1 protein expression with a threshold of 152 calculated from Cutoff Finder. Additionally, there were 46.5% of patients with TP53 mutation and 16.3% of patients with EGFR mutation.

## METHOD DETAILS

The overview of SRS is outlined in Figure 1. It consisted of three separate cascaded CNN modules: i) a tumor search module (TSM) to perform subtype lesion detection for pinpointing histologic adenocarcinoma areas and identifying corresponding tumor subtypes on H&E WSIs, ii)a ROB search module (RSM) to accomplish out-of-distribution (OOD) detection for selecting the most salient ROBs from predicted subtype lesions and iii) a status prediction module (SPM) for predicting target biomarker status for each saliency ROB. In addition, we clarified the theoretical details of the framework in the Supplementary Method.

### Tumor search module

To alleviate the computation burden, we used fixed-sized windows to generate available-sized input from down-sampled WSIs equivalent to 5× object lens magnification (2.0 μm/pixel) for subtype lesion detection. Specifically, we set the window size to 1000 × 1000 pixels (2000 × 2000 μm), ensuring the range was large enough to contain multiple subtype lesions. During training, we placed windows centered on each annotated subtype lesion to generate tiles as training data. At the inference stage, we continued to use sliding windows of size 1000 × 1000 pixels (2000 × 2000 μm) to partition WSIs with an overlap of 250 pixels (500 × 500 μm) (see Figure S5). The overlap ensured that any lesion would be entirely contained in some window. Additionally, we discarded the sliding-window-generated tiles with little tissue (i.e., tiles with >50% areas consisting of background pixels with each RGB value >220). However, the existence of overlap could cause the redundancy of predictions. Therefore, we perform non-maximum suppression (NMS) twice to reduce redundancy. The first NMS was applied after inferring subtype regions on each sliding-window-generated tile, and the second one was applied after mapping predicted bounding boxes back to the original WSI. By applying NMS twice, we were able to reduce redundancy and improve the efficiency and accuracy of detection.

In TSM, we utilized Cascade R-CNN[34] as a basic detector to detect multiple histologic subtypes of adenocarcinomas on WSIs, which contains a set of R-CNN detectors. The architecture had an advantage in improving the quality of predicted subtype locations stage by stage with respect to the prior dominant object detection model named Faster R-CNN.[35] It consisted of three modules, namely backbone, neck and head. The Res2Net[36] served as backbone to extract semantic feature from pathological images. It represented multi-scale features at a granular level and enlarged the size of receptive fields. The neck was Feature Pyramid Network (FPN),[37] constructing features pyramid with marginal extra cost on computation burden and memory storage to improve the detection accuracy. In this study, we merely utilized the feature output from the last three layers in ResNet as input to the FPN. The head consisted of the region proposal network (RPN)[35] and the R-CNN. The RPN scanned the feature map to output a set of anchor-based object proposals, each with a probability of foreground and background. Those anchors with different scales and aspect ratios were introduced to process objects with various shapes, whose design were based on prior knowledge. We employed a data-centric search named k-means clustering[38] to determine the optimal anchor design based on the characteristics of the annotated bounding boxes in the training set. As a result, we set a scale of 8 and aspect ratio of 0.8, 1.0 and 1.25 to generate anchors on each feature map in FPN. Additionally, the R-CNN resampled the object proposals generated by RPN with progressively increasing IoU thresholds to screen higher quality proposals and performed bounding box regression and object classification for each proposal. To alleviate the problem of imbalance between positive and negative samples, we used online hard example mining (OHEM) sampler[39] at each stage of RCNN. The OHEM sampler ranked the proposals by loss and only made use of the current worst-performing proposals to further optimize the network parameters.

We exemplified the detection performance of TSM by using LUAD samples in the TCGA, NLST, and WHTJ benchmark datasets. The lesion detectors were trained and tested independently on each dataset. For detecting five LUAD tumor subtype areas in a WSI, with an IoU threshold of 0.5, TSM achieved test mAP of 0.696, 0.738, and 0.712 for TCGA, NLST, and WHTJ, with recalls of 0.893, 0.923, and 0.820, respectively. We reported the detailed category-wise detection performance in Table S2.

To produce the appropriate data for the sequential RSM module, we enlarged the predicted bounding box to a square with a side length of N×512 pixels at an object lens magnification of 20×. This also allowed us to capture more paracanerous tissue with a clinical value around the tumor areas. The post-processing step is illustrated in Figure S6. The scale factor of $N$ was chosen to be the smallest integer making N×512 greater than the longer side of the predicted bounding box. Next, we divided the enlarged lesion bounding box into a set of 512 × 512 pixels grids. During the evaluation of the detection performance, we never performed the post-processing step.

## ROB search module

To enhance the accuracy of ROB positioning, we applied an out-of-distribution detection method based on confidence estimation.[40] This can help to identify the most valuable lesions for a target biomarker, referred to as saliency ROBs, from the subtype lesions predicted by the TSM. By doing so, we were able to eliminate the uninformative or irrelevant regions, which facilitates the model training of biomarker status classification and the ensemble of prediction on local tissue regions.

The ROB filter used in RSM was a Light-Xception architecture, equipped with an extra confidence branch paralleling the classification branch, referred to as Light-Xception-CE (see Figure S7), with which we can estimate the confidence score for a given subtype lesion. The Light-Xception was entirely based on separable convolution layers instead of conventional convolution. It was essentially a simplified variant of Xception.[41] Due to the decrease of model complexity, the Light-Xception had a faster speed to operate an image and a better ability to avoid overfitting. The separable convolution layer consisted in a depth-wise convolution layer that performs spatial convolution on feature map over each channel independently, followed by a point-wise convolution layer to perform affine transformation on full channel outputs over each spatial point. Compared with the conventional convolution, separable convolution had fewer trainable parameters and less computation with the same kernel size. The Light-Xception had 13 convolution layers, including 12 separable convolution layers. Those convolution layers were structured into 7 modules, and there are residual connections between adjacent modules, except for the first and the last module. The prediction branch outputed a normalized score between 0 and 1 with the softmax function, representing the probabilities of behaving high and low biomarker status. The confidence branch outputed a normalized

score between 0 and 1 with the sigmoid function, indicating how sure the model was to produce a correct prediction. We set a threshold for the confidence score and chose the subtype lesions with scores above the preset threshold as saliency ROBs. These ROBs were considered to have the greatest contribution to the model's decision-making process for biomarker status.

During training, we used the biomarker status of the source patient as the supervisory signal for each lesion. The RSM was optimized by integrating a confidence score output with a cross-entropy loss for the half samples in a mini-batch. This enabled the RSM to tolerate the ambiguous lesions exhibiting uninformative to the target biomarker determination and the mismatched lesions exhibiting inconsistent biomarker status with the source patient. However, too many lesions with mismatched supervisory signals can harm the training process of the traditional classification problem. In this study, the optimal confidence threshold was determined based on the highest classification accuracy achieved by the Light-Xception-CE model on the ROB subset, meeting the threshold requirement from the training set. In our experiments, we set the confidence threshold to 0.5, 0.48, 0.55, 0.50 for TMB, PD-L1 protein expression, TP53 and EGFR mutations.

### Status prediction module

In SPM, we employed the same Light-Xception architecture to train the last biomarker classifier for predicting the probability of exhibiting a high/positive biomarker status for each saliency ROB. Similar to the training of RSM, we assigned the patients' biomarker status as the target supervision for their associated saliency ROBs when feeding them into the CNN model as input. This pseudo-label assignment was consistent with the standard deep-learning pipeline for WSI diagnosis as in previous works.[6,7,17] For instance, if a patient presented high TMB status, whose ROBs would be assigned as high TMB status during CNN training. Summarily, we attempted to create a new representation for patients that replaces the original WSI while containing fewer meaningless patches that could interfere with the biomarker status decision.

### Prediction ensemble

We described how to aggregate the local patch predictions to derive the ultima biomarker status for a patient. We collected predictions of saliency ROBs from all WSIs associated with the source patient who may have multiple slides and took the median of patch predictions as the resulting diagnosis about the target biomarker.

### Experimental setup

We trained the Cascade R-CNN detector (TSM) based on the implementation of mmDetection.[42] All cascade detection stages in Cascade R-CNN had the same architecture, three stages for detection with the increasing IoU threshold of 0.3, 0.4, 0.5. The standard horizontal/vertical flipping and color jittering (including brightness, contrast, saturation, and hue) were applied as data augmentation. Furthermore, we randomly adjusted the width and height of bounding boxes of ground truth at a small range to alleviate the problem of boundary blurring. We utilized SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4 to update the network parameters via backpropagation. The training started on two synchronized GPUs, each holding 2 images per iteration, with learning rate of 0.001 using warm-up at the first 1000 iterations to linearly increase the learning rate from 1e-6. The learning rate was reduced by a factor of 10 at the 15th, 25th and 35th epoch and the training terminated at the 40th epoch.

We implemented the Light-Xception-CE (RSM) and Light-Xception (SPM) models with PyTorch. Along with the data augmentation mentioned during the TSM training process, we randomly adjusted the amount of hematoxylin and eosin stained by decomposing the tissue image from the RGB color space into the HED color space. We also applied Cutout[43] to randomly mask parts of the tissue regions. Additionally, we utilized the confidence loss function defined in the confidence estimation method[40] and SGD optimizer with a momentum of 0.95, weight decay of 5e-4, learning rate of 0.01 and a batch size of 64 to update the parameters of the Light-Xception-CE via backpropagation for 100 epochs. To address data imbalance, we set the loss weight to 1:2, 1:4, 1:1, 1:4 for biomarkers of TMB, PD-L1 protein expression, TP53, and EGFR mutations, respectively. Subsequently, we utilized the cross-entropy loss and SGD optimizer with a momentum of 0.95, weight decay of 5e-4, learning rate of 0.01 and a batch size of 32 to update the parameters of the Light-Xception via backpropagation for 100 epochs. The loss weight was set to 1:2, 1:3 and 1:1, 1:4 for biomarkers of TMB, PD-L1 protein expression, TP53, and EGRF mutations, respectively.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Model performance evaluation

The dataset splits for the TCGA, NLST, and WHTJ datasets were performed at the patient level using stratified random permutation. The SRS performance was validated on the TCGA test set, and the CPTAC dataset served as an external dataset. We evaluated the model performance based on the metrics, including sensitivity, specificity, average precision (AP), and area under the curve (AUC). For every possible probability cutoff, the ROC curve summarized the trade-off between sensitivity and specificity, while the PR curve summarized the trade-off between precision and recall. To assess the statistical significance of the AUC/AP score, we estimated the 95% confidence interval (CI) with bootstrap resampling, and derived the p value by comparing the observed AUC/AP score with the distribution of bootstrap AUC/AP scores. We based the hyperparameter and model developments on the validation split of the TCGA dataset with grid search.