Symb-xMIL: Symbolic Explanations for Multiple Instance Learning in Digital Pathology

Yanqing Luo, Julius Hense, Niklas Prenißl, Andreas Mock, Klaus-Robert Müller, Thomas Schnake*, Mina Jamshidi Idaji*

BIFOLD · TECHNISCHE UNIVERSITÄT BERLIN · LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN · UNIVERSITY OF TORONTO · CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN · NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

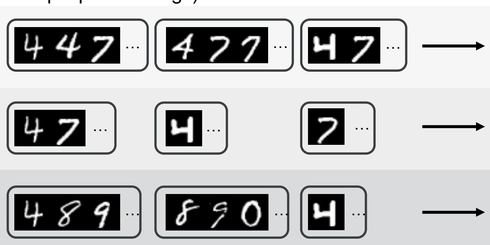
Going beyond *"where the model looks"* to *"how the model reasons"* across semantic features

Current MIL explanations highlight influential regions, but they **cannot** explain:
- **How** evidence is combined across regions
- Whether decisions rely on **morphologically meaningful features**
- How model strategies differ **across patient cohorts**

Symbolic xMIL maps model behavior to **human-readable** logical rules over **semantic features,** such as: `Tumor OR (V) Stroma, Normal → Mix`. This enables both per-slide interpretation and cohort-level analysis.

Validation on Synthetic Datasets

Ground truth (example positive bags)



Symb-xMIL result

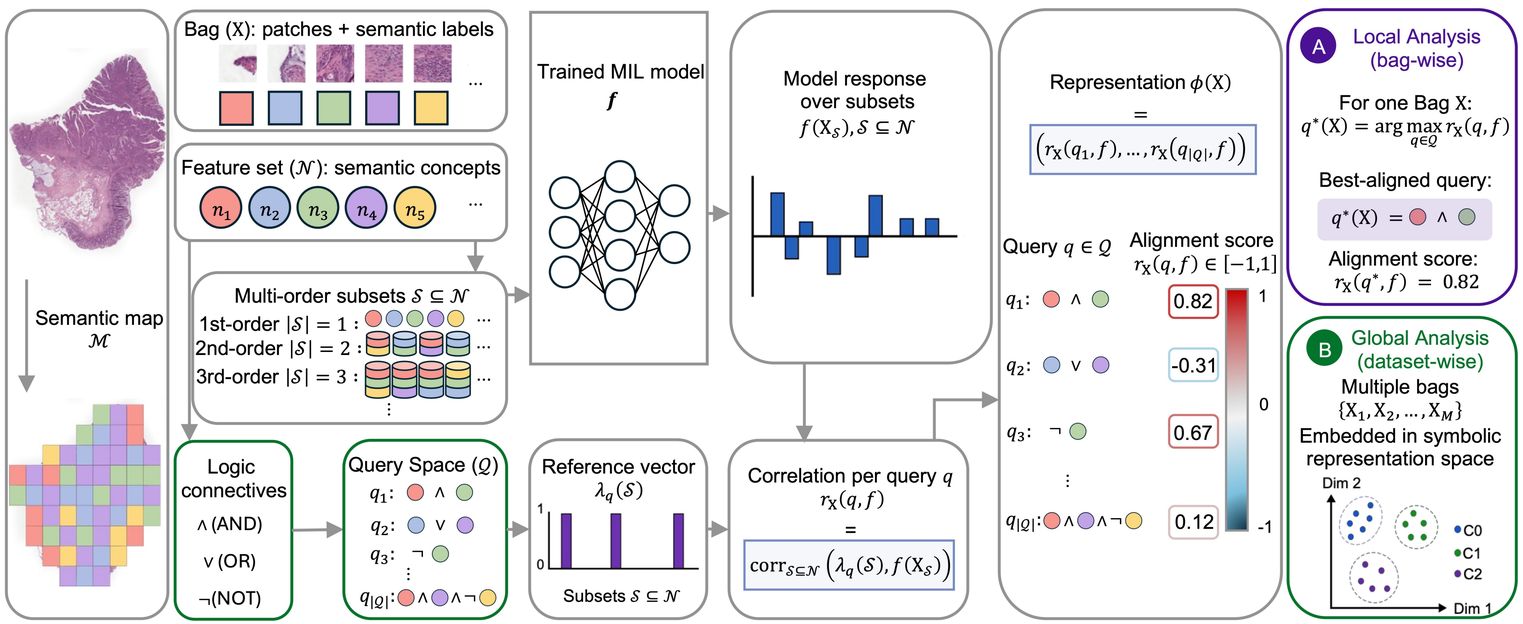
Query	Mean alignment
4 ∧ 7	**0.951**
4 ∧ (7 ∨ R)	0.761
4	0.658

Query	Mean alignment
4 ∨ 7	**0.992**
4 ∨ ¬R	0.675
7 ∨ ¬R	0.669

Query	Mean alignment
4 → 7	**0.967**
7	0.779
7 ∨ R	0.659

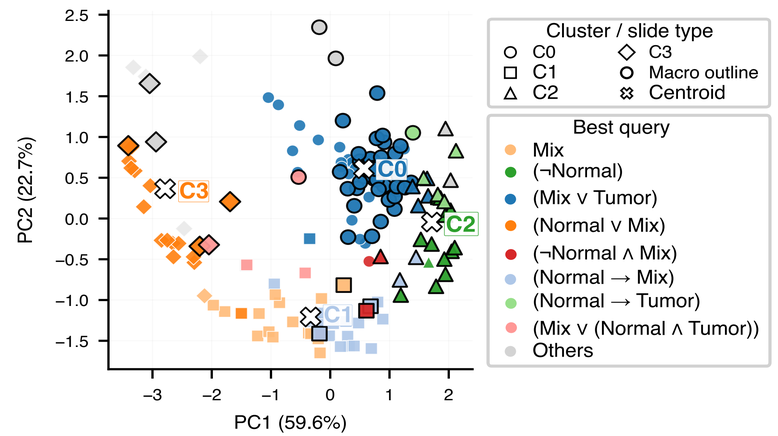
Method

1 Input & Pre-processing · 2 Semantic features; Queries & subsets · 3 Model evaluation on Subsets & reference vector · 4 Similarity to logical queries · 5 Symbolic representation space



Validation on Camelyon16

We trained a TransMIL classifier on Camelyon16 dataset for tumor classification task. **Test AUROC = 0.995** — nearly perfect. Symbolic xMIL reveals **4 distinct reasoning clusters** among tumor slides:



Cluster	Dominated rule	Interpretation	N
C0	Mix ∨ Tumor	**Expected reasoning**: tumor and adjacent mixed regions drive the prediction.	68
C1	Normal → Mix	**Micrometastasis** pattern: mixed regions provide positive evidence, while normal tissue alone is insufficient.	41
C2	¬Normal	Strict tumor-driven reasoning: normal tissue provides negative evidence, mostly observed in **macrometastases**.	26
C3	Normal ∨ Mix	**Potential shortcut**: normal patches alone can trigger tumor-positive predictions.	25

If C3's normal patches really were "just normal," transferring them into ground-truth normal slides should change nothing. We tested this, **21 / 25** transferred slides flipped to tumor-positive (84%)

Symbolic xMIL exposes a hidden error behind a near-perfect model

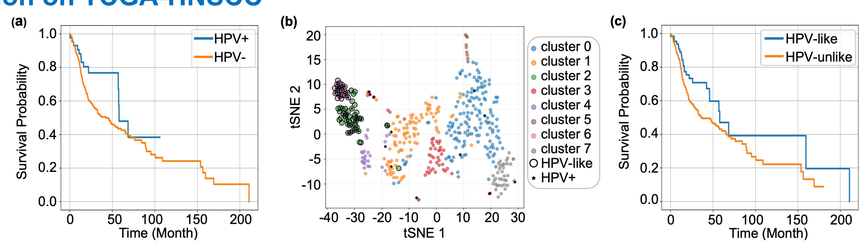
Validation on TCGA-HNSCC

Task: predict HPV status from H&E slides

Question: does the *symbolic representation* of an HPV-prediction model encode survival information **beyond the binary HPV label itself**?

Approach: cluster patients in symbolic representation space.
- Clusters enriched for HPV+ → label as **HPV-like**;
- others → **HPV-unlike**. (Some HPV+ patients land in HPV-unlike cluster)

Hypothesis-generation: the symbolic representation space captures biologically meaningful patterns, extending beyond binary HPV status. Worth investigating in larger cohorts.



Clinical baseline: KM analysis logrank p=0.06

Symbolic representation clusters

Symbolic-space-drived partition: KM analysis logrank p =0.04

Take-home

- Beyond heatmap-based explanation: explanations as logical rules could capture higher-order interactions
- Local explanations reveal hidden errors; global cohort analysis stratifies survival subgroups

Relevant Work:
[1] Schnake et al. „Towards symbolic XAI—explanation through human understandable logical relationships between features.“ In: Information Fusion
[2] Idaji, Mina Jamshidi, et al. "Beyond attention heatmaps: How to get better explanations for multiple instance learning models in histopathology." Medical Image Analysis
[3] Hense, Julius, et al. "xMIL: Insightful explanations for multiple instance learning in histopathology." Advances in Neural Information Processing Systems 37