Biomarker Learning from Whole Slide Images: A Comprehensive Review from Hand-Crafted Features to Foundation Models

Wassim Chikhi* and Hadia Amjad*
*Université Paris Cité, M2 Vision Multimédia et Interactions
Supervisors: Pr. Nicole Vincent, Pr. Nicolas Loménie, Pr. Camille Kurtz
Email: {wassim.chikhi, hadia.amjad}@etu.u-paris.fr

***Abstract*—Whole Slide Images (WSI) have revolutionized computational pathology by enabling gigapixel-scale digital analysis of histological tissue sections. However, their extreme scale, weak slide-level supervision, and substantial cross-site domain shift pose fundamental challenges for automated biomarker learning. This paper provides a comprehensive review of computational methods for predicting diagnostic, prognostic, and molecular biomarkers directly from H&E-stained WSIs. We organize the literature around three key questions: (i) How are patch-level representations learned?—from hand-crafted texture descriptors through CNN-based classification to self-supervised pretraining and recent pathology foundation models (UNI, Virchow, H-Optimus-0); (ii) How are patch embeddings aggregated into slide-level predictions?—spanning heuristic pooling, attention-based Multiple Instance Learning (MIL), Transformer architectures, and graph neural networks; (iii) How are biomarkers rigorously evaluated?—covering patient-level data splitting, cross-site validation, calibration metrics, and few-shot learning protocols. We synthesize findings from over 30 seminal works, identify five persistent challenges (dense segmentation limitations, loss of spatial structure, data bias, interpretability gaps, computational accessibility), and discuss promising future directions including WSI-level self-supervised learning, hybrid CNN-ViT architectures, causal modeling, and federated learning approaches.**

***Index Terms*—Computational pathology, whole slide images, multiple instance learning, self-supervised learning, foundation models, biomarker prediction, deep learning, histopathology.**

I. Introduction

A. Context and Motivation

Digital pathology has fundamentally transformed tissue analysis through Whole Slide Images (WSIs)—high-resolution digital scans of H&E-stained glass slides captured at diagnostic magnifications (20× or 40×). A single WSI can exceed 100,000 × 100,000 pixels at native resolution, representing several gigabytes of data and containing billions of pixels across diverse tissue compartments: epithelium, stroma, immune infiltrate, vasculature, and necrotic regions.

From these images, pathologists derive critical clinical endpoints that guide patient management:

- **Diagnostic labels**: Tumor presence, histological subtype (adenocarcinoma vs. squamous cell carcinoma), benign vs. malignant classification
- **Grading scores**: ISUP grades 1–5 for prostate cancer (based on Gleason patterns), Nottingham scores for breast cancer (tubule formation, nuclear pleomorphism, mitotic activity)
- **Molecular surrogates**: Microsatellite instability (MSI-H vs. MSS) [14], oncogenic mutations (KRAS, EGFR, TP53, BRAF) [15], homologous recombination deficiency (HRD), transcriptomic subtypes (CMS in colorectal cancer [14])

Computational biomarker learning aims to predict these endpoints directly from WSI morphology, offering potential clinical advantages: reduced cost (avoiding molecular assays), faster turnaround times, applicability to archival tissue lacking genomic profiling, and retrospective analysis of large patient cohorts.

B. Three Fundamental Challenges

WSI-based biomarker prediction confronts three interrelated computational obstacles:

1. Extreme scale and memory constraints. Native-resolution WSIs (e.g., 100,000 × 100,000 pixels at 40× magnification) cannot fit in GPU memory. Standard pipelines decompose slides into thousands of fixed-size patches (256×256 or 512×512 pixels), introducing a two-stage paradigm: patch-level feature extraction followed by slide-level aggregation. This tiling process can yield 10,000–50,000 patches per slide, each requiring independent encoding.

2. Weak supervision and label scarcity. Clinical labels are available at slide or patient level, not for individual patches or pixels. This precludes standard supervised learning and necessitates weakly supervised frameworks—most prominently Multiple Instance Learning (MIL) [5], [6]—where the slide is modeled as a "bag" of unlabeled patch "instances." Moreover, obtaining expert annotations is expensive and time-consuming, resulting in small labeled datasets relative to the model complexity.

3. Domain shift and generalization. Staining protocols (hematoxylin concentration, eosin penetration time), scanner hardware (Aperio vs. Hamamatsu vs. Leica), tissue preparation

(FFPE fixation duration, section thickness), and patient demographics vary substantially across hospitals [1]. Models trained on single-center cohorts often exhibit degraded performance when deployed to new sites, limiting clinical translation.

C. Historical Progression and Scope

The field has evolved through distinct methodological eras:

1) **Hand-crafted features** (pre-2015): Haralick texture descriptors [4], Local Binary Patterns, color histograms, morphometric features combined with classical ML (SVM, Random Forest)
2) **CNN-based patch classification** (2015–2018): ImageNet-pretrained ResNets with heuristic aggregation (mean/max pooling)
3) **Attention-based MIL** (2018–2021): Learned weighted aggregation of patch embeddings [5], instance clustering [6]
4) **Self-supervised learning** (2020–2023): Contrastive methods (SimCLR [7], MoCo [8]) and distillation-based approaches (DINO [9], DINOv2 [10]) adapted to histology [11]
5) **Foundation models** (2023–present): Large-scale vision encoders (UNI [12], Virchow, Prov-GigaPath) and multimodal architectures (H-Optimus-0 [13]) pretrained on 100k+ WSIs

This review synthesizes findings from these eras, organizing the literature around three axes:

- **Section II–IV**: Patch representation learning (hand-crafted → CNN → SSL → foundation models)
- **Section V**: Slide-level aggregation strategies (pooling → MIL → Transformers → graphs)
- **Section VI**: Biomarker types, datasets, and evaluation protocols
- **Section VII**: Persistent challenges and future directions

Acronyms. WSI (Whole Slide Image), ROI (Region Of Interest), MIL (Multiple Instance Learning), SSL (Self-Supervised Learning), ViT (Vision Transformer), ISUP (International Society of Urological Pathology), MSI (Microsatellite Instability), HRD (Homologous Recombination Deficiency), TCGA (The Cancer Genome Atlas).

II. WSI Properties and Preprocessing

A. Multi-Resolution Pyramid Structure

WSIs are stored as multi-resolution pyramids with 3–5 magnification levels (e.g., 40×, 20×, 10×, 5×, 1.25×). Tissue segmentation is performed at low magnification (1–2×) to exclude background regions (glass, pen marks, coverslip artifacts), then foreground regions are tiled into patches at diagnostic magnification (typically 20× or 40×).

B. Standard Preprocessing Pipeline

The canonical WSI processing workflow comprises four stages:

1. Tissue detection. Otsu thresholding or morphological operations applied to downsampled images (1× or 2× magnification) separate foreground tissue from background glass.

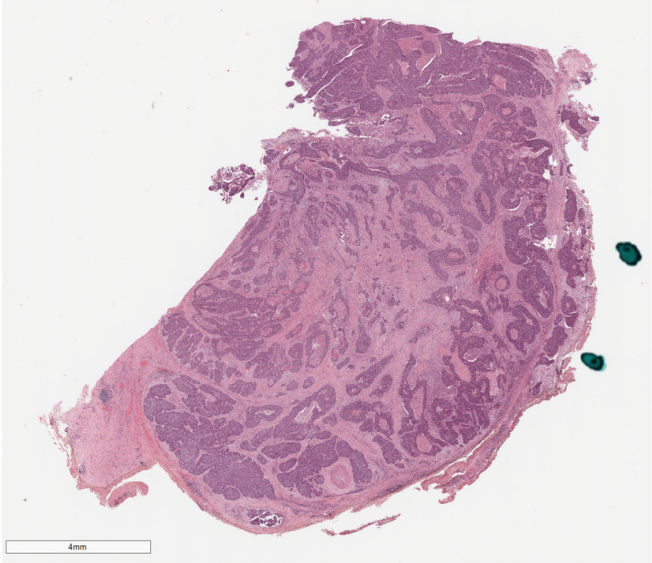


Fig. 1. Multi-scale WSI structure. Low magnification (1–2×) reveals global tissue architecture; intermediate levels (5–10×) show regional organization (glands, stroma); high magnification (20–40×) captures cellular morphology (nuclei, cytoplasm, extracellular matrix).

2. Tiling. Foreground regions are decomposed into non-overlapping or overlapping patches (common sizes: 256×256, 512×512 pixels at 20× or 40× magnification). A typical slide at 20× yields 2,000–10,000 patches; at 40×, this can exceed 50,000.

3. Quality filtering. Patches are discarded based on:

- Low tissue content (<50% foreground pixels)
- Excessive blurring (Laplacian variance below threshold)
- Color artifacts (saturation/brightness outside expected ranges)

4. Color normalization (optional). To mitigate staining variation across scanners and laboratories, several normalization techniques have been proposed:

- **Macenko method** [2]: Stain deconvolution via singular value decomposition, separating hematoxylin and eosin channels, then rescaling to reference statistics
- **Vahadane method** [3]: Structure-preserving color normalization via sparse non-negative matrix factorization

Tellez et al. [1] quantified the impact of these preprocessing choices, demonstrating that stain normalization can reduce domain shift effects but also risks overfitting to training-set color distributions. The trade-off between normalization benefits and potential information loss remains task-dependent.

C. Biomarker Categories

WSI-derived biomarkers span three classes:

Morphological biomarkers quantify tissue architecture directly observable by pathologists:

- Tumor-infiltrating lymphocyte (TIL) density [16]
- Necrosis extent and spatial distribution
- Glandular disorganization (cribriform patterns, loss of polarity)

- Nuclear pleomorphism, mitotic count, apoptotic index

Grading systems assign ordinal scores based on morphological patterns:

- ISUP grading for prostate cancer: grades 1–5 derived from Gleason patterns (well-formed glands to solid sheets)
- Nottingham grading for breast cancer: tubule formation, nuclear atypia, mitotic activity

Molecular surrogates aim to predict genomic or transcriptomic features without sequencing. Kather et al. [14] demonstrated pan-cancer detection of microsatellite instability (MSI) and clinically actionable genetic alterations directly from H&E morphology, achieving AUROCs of 0.77–0.84 across six cancer types. Coudray et al. [15] predicted STK11, EGFR, and TP53 mutations in non-small cell lung cancer (NSCLC) from histology images with AUROCs exceeding 0.70.

However, the clinical reliability and utility of morphology-only molecular predictions remain debated. Challenges include:

- Correlation does not imply causation: morphological patterns may reflect downstream effects rather than direct genomic drivers
- Generalization to rare mutations and heterogeneous tumors is uncertain
- Regulatory approval requires rigorous multi-site validation, which is often lacking

III. Patch Representation Learning: Hand-Crafted to Deep Features

A. Hand-Crafted Feature Era (Pre-2015)

Early computational pathology relied on domain-specific feature engineering, combining insights from image processing and pathology expertise:

Texture descriptors:

- **Haralick features** [4]: Statistical measures derived from gray-level co-occurrence matrices (GLCM)—contrast, correlation, energy, homogeneity—capturing spatial relationships between pixel intensities
- **Local Binary Patterns (LBP)**: Rotation-invariant texture codes encoding local intensity gradients
- **Gabor filter banks**: Multi-scale, oriented convolutions extracting edge responses at different frequencies and angles

Color features:

- RGB, HSV, or LAB color histograms summarizing staining intensity distributions
- Stain deconvolution separating hematoxylin (nuclear) and eosin (cytoplasmic) channels [2]

Morphometric features:

- Nuclear shape descriptors (area, perimeter, circularity, eccentricity, solidity)
- Gland topology via Voronoi diagrams, Delaunay triangulation, or minimum spanning trees
- Spatial statistics: nearest-neighbor distances, clustering indices

These features were combined and fed to classical machine learning classifiers (Support Vector Machines, Random Forests, Gradient Boosting). While interpretable and aligned with pathologists' reasoning, they suffered from:

- Task-specific tuning: feature sets required manual re-engineering for each tissue type, staining protocol, or clinical endpoint
- Poor robustness to color variation: scanner-to-scanner differences degraded performance
- Limited scalability: computing these features for millions of patches is computationally expensive
- Inability to capture hierarchical tissue structure: hand-crafted descriptors operate at fixed scales and miss multi-level organization

B. CNN-Based Patch Classification (2015–2018)

Deep convolutional neural networks (CNNs) revolutionized patch-level representation learning by automatically discovering hierarchical features from raw pixels. The standard paradigm:

1) Pretrain a CNN (typically ResNet-50, VGG-16, or Inception-v3) on ImageNet (1.2M natural images, 1000 classes)
2) Fine-tune on labeled histology patches (when available) or use as a frozen feature extractor
3) Aggregate patch-level predictions via heuristics: mean pooling, max pooling, majority vote, or top-k selection

This approach yielded substantial improvements over hand-crafted features but had critical limitations:

Label scarcity. Patch-level annotations are rare in clinical datasets. Even when available (e.g., tumor vs. normal labels for small ROIs), they require expert pathologist time and may exhibit inter-observer variability.

Label noise. A common workaround assigns the slide label to all patches within that slide. However, this is a noisy assumption: a slide labeled "tumor" may contain abundant normal stroma, blood vessels, or lymphocytes. Training on such noisy labels dilutes the signal and can degrade performance.

Domain gap. ImageNet-pretrained features are optimized for natural images (objects, animals, scenes) and may not capture histology-specific textures (fibrous stroma, glandular lumens, inflammatory infiltrates) or color distributions (H&E staining vs. RGB photography).

C. Multiple Instance Learning (MIL) Framework

MIL provides a principled formalization of weak supervision tailored to WSI analysis [5], [6]. Each WSI is treated as a *bag* $\mathcal{B} = \{x_1, \ldots, x_n\}$ of patch instances, where only the bag label $y_{\text{slide}} \in \{0, 1\}$ is observed. The standard MIL assumption: a bag is positive if *at least one* instance is positive, and negative if *all* instances are negative.

1) Attention-Based MIL: Ilse et al. [5] introduced a neural attention mechanism for MIL that learns instance-specific importance weights. The architecture comprises:

1. Instance encoding. A shared encoder f_θ (CNN or ViT) maps each patch to an embedding:

$$h_i = f_\theta(x_i) \in \mathbb{R}^d, \quad i = 1, \ldots, n \tag{1}$$

2. Attention scoring. A small learnable network computes attention weights:

$$a_i = w^\top \tanh(V h_i) \in \mathbb{R} \tag{2}$$

$$\alpha_i = \frac{\exp(a_i)}{\sum_{j=1}^{n} \exp(a_j)} \tag{3}$$

where $V \in \mathbb{R}^{d \times d}$ and $w \in \mathbb{R}^d$ are trainable parameters. The attention weights α_i reflect the model's belief that patch i is informative for the slide label.

3. Bag aggregation. The slide representation is a weighted sum:

$$H_{\text{slide}} = \sum_{i=1}^{n} \alpha_i h_i \in \mathbb{R}^d \tag{4}$$

4. Classification. A final MLP g_ϕ predicts the slide label:

$$\hat{y}_{\text{slide}} = g_\phi(H_{\text{slide}}) \tag{5}$$

This framework has several advantages: (i) end-to-end trainable with slide-level labels only; (ii) attention weights provide interpretability (highlighting informative patches); (iii) compatible with any patch encoder f_θ.

2) CLAM: Clustering-Constrained Attention MIL: Lu et al. [6] extended attention-based MIL with two key innovations:

1. Instance clustering. Patch embeddings are clustered (via k-means) into putative positive and negative groups. This encourages the model to discover diverse morphological patterns rather than overfitting to a single representative patch.

2. Hard negative sampling. During training, CLAM samples hard negatives (patches with high attention scores in negative bags) and hard positives (low-attention patches in positive bags) to sharpen decision boundaries.

On TCGA benchmarks, CLAM achieved 3–8% AUC improvement over standard attention-based MIL [6], particularly on tasks with high intra-slide heterogeneity (e.g., distinguishing tumor subtypes with overlapping morphologies).

MIL has become the *de facto* standard for slide-level prediction, even when the encoder f_θ is replaced by large foundation models (Section IV).

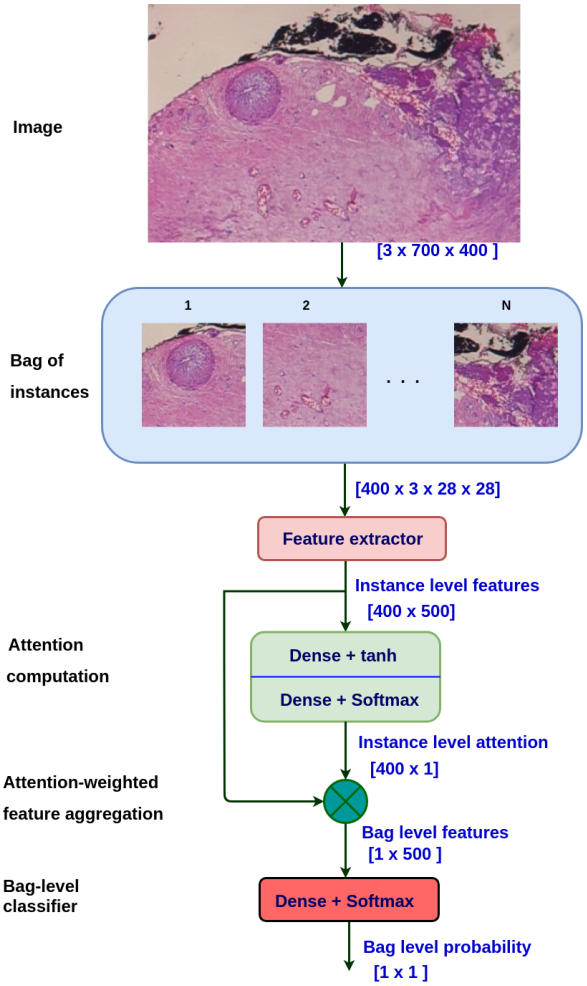


Fig. 2. Attention-based MIL architecture [5]. Patch embeddings h_i are weighted by learned attention scores α_i and aggregated into slide representation H_{slide} for classification.

IV. Self-Supervised Learning for Histopathology

A. Motivation: Leveraging Unlabeled Data

Hospitals accumulate vast archives of unlabeled WSIs (millions of slides across diverse organs and pathologies), but obtaining expert annotations is expensive, time-consuming, and subject to inter-observer variability. Self-supervised learning (SSL) offers a paradigm to exploit this unlabeled data by training models on pretext tasks—predicting one augmented view of a patch from another, clustering similar patches, or reconstructing masked regions—without manual labels.

SSL is particularly appealing for histopathology because:

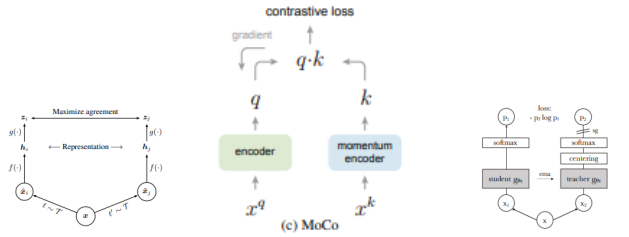
- **Domain adaptation**: Pretraining on histology patches reduces the domain gap relative to ImageNet weights
- **Robustness to domain shift**: SSL models trained on diverse multi-site cohorts generalize better to new scanners and staining protocols [1]
- **Few-shot learning**: Rich pretrained representations enable strong performance with very few labeled examples (Section VI-D)

B. Contrastive Self-Supervised Learning

Contrastive methods learn representations by maximizing agreement between different augmented views of the same patch (positive pairs) while pushing apart embeddings of different patches (negative pairs).

1) SimCLR: Simple Framework for Contrastive Learning: Chen et al. [7] introduced SimCLR, which operates as follows:

1) Sample a mini-batch of N patches
2) Apply two random augmentations to each patch, yielding $2N$ views
3) Encode each view with a shared CNN backbone, followed by a projection head (MLP)
4) Maximize cosine similarity between augmentations of the same patch, minimize similarity to all other patches in the batch



(a) SimCLR (b) MoCo (c) DINO

Fig. 3. Self-supervised learning paradigms: (a) SimCLR maximizes agreement between augmented views in a single forward pass with large batches; (b) MoCo uses a momentum encoder and queue of negatives; (c) DINO employs student-teacher distillation with multi-crop training and centering to avoid collapse.

The loss for a positive pair (i, j) is:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{6}$$

where z_i, z_j are ℓ_2-normalized projections, $\text{sim}(\cdot, \cdot)$ is cosine similarity, and τ is a temperature hyperparameter.

SimCLR requires large batch sizes ($N \geq 256$) to provide sufficient negatives for effective contrastive learning, necessitating multi-GPU training.

2) MoCo: Momentum Contrast: He et al. [8] addressed the batch size limitation via a momentum-updated encoder and a queue of negative examples. Key innovations:

- **Queue**: Maintains a large dictionary of negative patch embeddings (e.g., 65,536) decoupled from batch size
- **Momentum encoder**: Updates the encoder for negative keys as an exponential moving average (EMA) of the query encoder, ensuring consistency of queue representations across training steps

MoCo achieved comparable or superior performance to SimCLR with smaller batches, reducing hardware requirements.

3) CTransPath: Contrastive Learning with Vision Transformers: Wang et al. [11] adapted contrastive SSL to histopathology using a Swin Transformer backbone (86M parameters). Pretrained on 15 million patches from TCGA and PAIP datasets spanning multiple cancer types, CTransPath demonstrated:

- 3–7% AUC improvement over ImageNet-pretrained ResNet-50 on slide-level classification tasks (tumor vs. normal, subtype prediction)
- Strong cross-site generalization: models trained on TCGA slides transferred well to independent hospital cohorts
- Superior performance on rare cancer types with limited labeled data

This work was among the first to show that Vision Transformers (ViTs) pretrained on histology patches can outperform CNNs, paving the way for larger foundation models.

C. Distillation-Based Self-Supervised Learning

Rather than contrasting positive and negative pairs, distillation-based methods use a student-teacher framework where the teacher provides soft targets for the student without requiring explicit negatives.

1) DINO: Self-Distillation with No Labels: Caron et al. [9] introduced DINO, which combines:

- **Student-teacher architecture**: Teacher network is an exponential moving average (EMA) of student weights
- **Multi-crop training**: Student sees multiple augmented crops of each image (2 global crops at 224×224, several local crops at 96×96); teacher sees only global crops
- **Cross-entropy objective**: Student is trained to match teacher's output distribution (after centering and sharpening transformations to prevent collapse)
- **No projection head discarding**: Unlike SimCLR, the backbone features themselves are used for downstream tasks

DINO demonstrated strong performance on ImageNet without labels and produced semantically meaningful attention maps highlighting object parts.

2) DINOv2: Scaling Up Distillation: Oquab et al. [10] scaled DINO to ViT-Large and ViT-Giant backbones (up to 1B parameters) trained on deduplicated ImageNet-22k and additional web-scraped data. Technical improvements include:

- ℓ_2-normalized projection heads and feature outputs
- Sinkhorn-Knopp centering of teacher outputs for stability
- Resolution-adaptive positional encodings
- Masked image modeling as an auxiliary task

When applied to histology patches, DINOv2 produces embeddings robust to color and geometric augmentations, transferring well to downstream MIL tasks with minimal fine-tuning. DINOv2 has become a popular initialization for pathology foundation models (Section IV).

V. Pathology Foundation Models

A. The Scaling Hypothesis

In natural language processing, scaling up model size and training data has led to dramatic improvements in performance and generalization (GPT-3, GPT-4). The same hypothesis is now being tested in computational pathology: if we pretrain very large vision encoders on hundreds of thousands or millions of WSIs, we should obtain broadly reusable representations that transfer well to many downstream tasks with minimal fine-tuning.

B. Vision-Centric Foundation Models

1) UNI: A General-Purpose Foundation Model: Chen et al. [12] introduced UNI, a ViT-Large encoder (307M parameters) pretrained with DINOv2 on approximately 100,000 WSIs from 20 medical centers, covering 17 major organs and over 40 cancer types. UNI was evaluated on 33 downstream tasks spanning:

- Slide-level classification (tumor vs. normal, subtype prediction)
- ROI-level classification (gland grading, immune cell detection)
- Patch retrieval (finding morphologically similar patches)

- Dense segmentation (tumor boundaries, tissue compartments)

Key findings:

- **Multi-task generalization**: UNI achieved state-of-the-art or near-state-of-the-art performance on 31 of 33 tasks, with particularly strong gains on rare cancers (up to +19% AUC vs. ImageNet-pretrained baselines)
- **Scaling trends**: Performance improved monotonically with pretraining data size (1k $\rightarrow$ 10k $\rightarrow$ 100k slides), yielding approximately +8% average AUC gain across tasks
- **Few-shot efficiency**: With only 4 labeled slides per class, UNI matched or exceeded the performance of ImageNet-pretrained ResNet-50 trained on 32 slides per class—an $8\times$ reduction in annotation requirements
- **Segmentation limitations**: Despite strong classification performance, dense segmentation tasks showed only modest improvements (ΔDice $\approx$ 0.005 vs. U-Net with ImageNet encoder), suggesting that ViT architectures optimized for global context may sacrifice fine-grained pixel-level localization

2) Virchow and Prov-GigaPath: Scaling Beyond 100k Slides: **Virchow** is a ViT-Huge model (632M parameters) pretrained on over 1.5 million WSIs via DINOv2. It exhibits similar trends to UNI—strong multi-task performance, excellent few-shot learning, cross-site robustness—with additional capacity benefits on complex tasks requiring nuanced morphological reasoning.

Prov-GigaPath is a hierarchical Vision Transformer trained on 1.3 billion tissue patches extracted from 171,000 WSIs. Unlike UNI and Virchow (which operate at patch level only), Prov-GigaPath incorporates both tile-level and slide-level pretraining objectives, aiming to capture multi-scale tissue organization.

C. Multimodal Foundation Models

1) H-Optimus-0: Vision + Genomics + Text: Saillard et al. [13] introduced H-Optimus-0, a multimodal foundation model jointly processing:

- **Vision**: WSI patch embeddings via a ViT encoder
- **Genomics**: Mutation profiles, copy-number alterations, gene expression vectors (RNAseq)
- **Text**: Free-text pathology reports (diagnosis, tumor stage, morphological descriptions) encoded via BERT-style language model

These three modalities are projected into a shared embedding space where cross-attention layers fuse information. The architecture is trained end-to-end on slide-level tasks (diagnosis, molecular biomarker prediction, survival outcome).

On molecular biomarker prediction (MSI status, HRD, tumor mutational burden), H-Optimus-0 achieves 5–12% AUC improvement over vision-only models. Attention maps indicate which modality contributed most to each prediction, providing interpretability.

However, multimodal models face practical challenges:

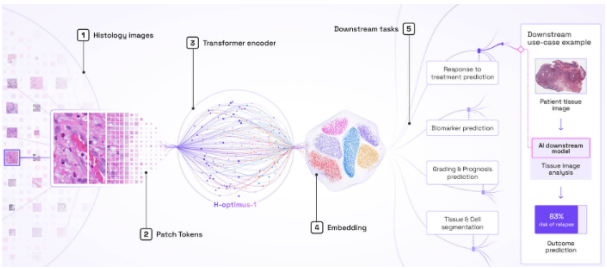


Fig. 4. Multimodal foundation model architecture (H-Optimus-0 [13]). Visual, genomic, and textual encoders project to a shared space; cross-attention layers enable modality fusion for downstream biomarker prediction.

- **Data requirements**: Require aligned WSI + genomics + reports, which are rare outside large research consortia (TCGA, institutional biobanks)
- **Training complexity**: Over 1 billion parameters, necessitating multi-node GPU clusters and sophisticated distributed training
- **Data curation overhead**: Matching WSIs to molecular assays across Electronic Health Record (EHR) systems is non-trivial and error-prone

D. Summary: Advantages and Limitations

Foundation models bring:

- **Multi-task reuse**: A single pretrained encoder can be fine-tuned or used zero-shot on many tasks (classification, retrieval, segmentation)
- **Label efficiency**: Strong few-shot performance ($<$10 slides/class) thanks to rich pretrained representations
- **Cross-site robustness**: Pretraining on diverse multi-institutional cohorts improves generalization to new scanners, staining protocols, and patient demographics

Persistent limitations:

- **Computational cost**: Training requires 100k+ WSIs, $8\times$ A100/H100 GPUs, and weeks of training time—limiting accessibility for smaller research groups
- **Data bias**: Training datasets overrepresent common cancers (lung, breast, colorectal) and H&E FFPE slides; rare cancers, special stains (PAS, Masson trichrome), cytology, and hematopathology are underrepresented
- **Segmentation weakness**: ViT architectures prioritize global context via self-attention but may sacrifice fine-grained spatial localization (ΔDice $\approx$ 0.005 vs. CNN baselines)
- **Interpretability gap**: High-dimensional embeddings ($d \approx$ 768–1024) don't map cleanly to pathologist-interpretable features (nuclear grade, mitotic count, cytoplasmic eosinophilia)

VI. Slide-Level Aggregation Strategies

Patch-level representations—whether from ResNet, ViT, SSL encoders, or foundation models—must be aggregated into slide-level predictions for biomarker learning. This section taxonomizes aggregation strategies.

A. Heuristic Pooling (Baseline)

The simplest approach applies fixed aggregation functions without learnable parameters:

Mean pooling:

$$H_{\text{slide}} = \frac{1}{n} \sum_{i=1}^{n} h_i \tag{7}$$

Max pooling (element-wise):

$$H_{\text{slide}} = \max_{i \in \{1,\ldots,n\}} h_i \tag{8}$$

Top-k pooling:

$$H_{\text{slide}} = \frac{1}{k} \sum_{i \in \text{Top-}k} h_i \tag{9}$$

where Top-k denotes the k patches with highest classifier scores.

Limitations: These methods are order-invariant, treat all patches uniformly (mean) or use only extreme values (max), and contain no learnable parameters to adapt to task-specific morphological patterns.

B. Attention-Based MIL

As described in Section III-C, attention mechanisms [5] learn instance-specific weights α_i to emphasize informative patches. Extensions include:

Gated attention [5]: Introduces a gating mechanism to modulate attention scores based on instance content.

Multi-head attention: Parallel attention branches capture different morphological patterns (e.g., one head for tumor nuclei, another for stroma, a third for immune infiltrate).

CLAM [6]: Combines instance clustering with hard negative sampling to sharpen decision boundaries.

C. Transformer-Based Aggregation

1) TransMIL: Full Self-Attention Over Patches: Shao et al. [17] applied a full Transformer encoder to the bag of patch embeddings:

$$H' = \text{Transformer}([h_1, \ldots, h_n]) \tag{10}$$

$$H_{\text{slide}} = \text{mean}(H') \quad \text{or} \quad H_{\text{slide}} = H'_{[\text{CLS}]} \tag{11}$$

Self-attention enables each patch to attend to all other patches, capturing long-range dependencies and contextual interactions (e.g., tumor-stroma interfaces, spatial gradients in immune infiltration). However, computational cost scales as $O(n^2)$ in the number of patches, limiting applicability to very large slides ($n > 10{,}000$).

TransMIL demonstrated 2–5% AUC improvement over attention-based MIL on several TCGA benchmarks, particularly on tasks requiring global tissue organization (e.g., distinguishing invasive vs. in-situ carcinoma).

D. Graph Neural Networks (GNNs)

Graph-based methods model spatial relationships explicitly via graphs $G = (V, E)$:

- **Nodes V:** Patches (represented by embeddings h_i)
- **Edges E:** Spatial proximity (e.g., k-nearest neighbors in coordinate space or embedding space)

Graph convolutions propagate information along edges:

$$h_i^{(\ell+1)} = \sigma\left(W^{(\ell)} h_i^{(\ell)} + \sum_{j \in \mathcal{N}(i)} \tilde{A}_{ij} h_j^{(\ell)} \right) \tag{12}$$

where $\mathcal{N}(i)$ denotes neighbors of node i, $\tilde{A}$ is a normalized adjacency matrix, and $W^{(\ell)}$ are learnable weights. After L layers, node embeddings are aggregated (via mean/max pooling or attention) into a slide representation.

Advantages: Explicit spatial priors enable modeling of tissue topology (tumor-stroma boundaries, immune infiltration fronts, vascular networks).

Limitations: Graph construction heuristics (choice of k, edge thresholds) are task-dependent and may not generalize across organs or pathologies. Noisy spatial layouts (e.g., tissue folds, artifacts) can degrade GNN performance.

VII. Biomarkers, Datasets, and Evaluation

A. Biomarker Types and Clinical Relevance

Morphological biomarkers quantify tissue architecture visible to pathologists:

- Tumor-infiltrating lymphocytes (TILs): Saltz et al. [16] demonstrated that TIL density quantified from H&E images correlates with immunotherapy response and survival outcomes across 13 cancer types
- Necrosis extent, stromal proportion, glandular disorganization
- Nuclear atypia, mitotic count (key components of grading systems)

Molecular surrogates predict genomic features from morphology alone:

- Microsatellite instability (MSI): Kather et al. [14] achieved AUROCs of 0.77–0.84 for MSI detection across six cancer types (colorectal, gastric, endometrial)
- Oncogenic mutations: Coudray et al. [15] predicted STK11, EGFR, and TP53 mutations in NSCLC with AUROCs >0.70
- Copy-number alterations, transcriptomic subtypes (CMS, PAM50)

However, the clinical utility of morphology-only molecular predictions remains uncertain. Challenges include limited generalization to rare mutations, lack of mechanistic understanding (correlation vs. causation), and regulatory barriers (FDA approval requires extensive multi-site validation).

B. Major Datasets and Benchmarks

The Cancer Genome Atlas (TCGA): The largest public repository of WSIs with matched genomic data. Contains ∼30,000 slides across 33 cancer types from over 11,000 patients. While invaluable for research, TCGA has limitations: retrospective curation, heterogeneous slide quality, scanner-specific color distributions.

CAMELYON16/17: Lymph node metastasis detection challenges with pixel-level annotations. Widely used for benchmarking preprocessing pipelines and aggregation methods, though not directly focused on biomarker prediction.

PAIP (Pathology AI Platform): Korean multi-institutional dataset for liver cancer (HCC) segmentation and survival prediction.

Institutional cohorts: Many foundation models are pretrained on private hospital datasets (e.g., UNI's 100k slides from 20 centers), which are not publicly accessible. This limits reproducibility and raises concerns about data bias.

C. Evaluation Protocols: Best Practices

1) Patient-Level Data Splitting: **Critical requirement:** All slides from a patient must belong to the same split (train/val/test) to avoid data leakage. Violating this can artificially inflate performance by 10–20% AUC [1].

Standard protocols:

- 5-fold or 10-fold cross-validation with patient-stratified folds
- Holdout split: 70% train, 15% val, 15% test (patient-level)
- External validation on independent cohorts (different hospitals, scanners) to assess cross-site generalization

2) Classification Metrics: **AUROC (Area Under ROC Curve):** Primary metric for binary and multi-class classification. Reports model's ability to rank positive instances higher than negatives. Should be accompanied by 95% confidence intervals via patient-level bootstrap (resample patients with replacement, $B \geq 1000$ iterations).

AUPRC (Area Under Precision-Recall Curve): Preferred for highly imbalanced datasets (e.g., rare cancer subtypes where prevalence $<5\%$).

Balanced accuracy: $\frac{1}{2}(\text{Sensitivity} + \text{Specificity})$ to account for class imbalance.

3) Segmentation Metrics: **Dice coefficient:**

$$\text{Dice} = \frac{2 \cdot |Y_{\text{pred}} \cap Y_{\text{true}}|}{|Y_{\text{pred}}| + |Y_{\text{true}}|} \tag{13}$$

Intersection over Union (IoU):

$$\text{IoU} = \frac{|Y_{\text{pred}} \cap Y_{\text{true}}|}{|Y_{\text{pred}} \cup Y_{\text{true}}|} \tag{14}$$

4) Few-Shot Learning Evaluation: Plot performance (AUROC, F1) as a function of labeled slides per class: 1, 2, 4, 8, 16, 32, 64. Foundation models typically show strong few-shot behavior, matching CNN baselines trained on 8× more data [12].

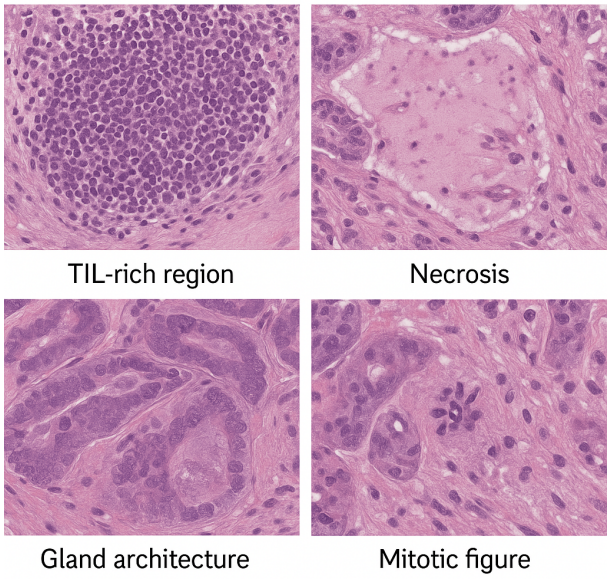


Fig. 5. Key morphological patterns for biomarker prediction: (a) TIL-rich immune infiltrate; (b) necrotic regions; (c) disrupted glandular architecture; (d) mitotic figures (arrows).

5) Calibration and Uncertainty: **Expected Calibration Error (ECE):** Measures alignment between predicted probabilities and true frequencies. Critical for clinical deployment where confidence scores guide triage decisions.

Reliability diagrams: Plot predicted probability bins vs. observed frequency.

VIII. Persistent Challenges and Future Directions

Despite rapid progress, five fundamental challenges remain unresolved.

A. Dense Segmentation Limitations

Foundation models based on ViTs (UNI, Virchow) show only marginal gains on pixel-level segmentation tasks (ΔDice $\approx$ 0.005 vs. U-Net with ImageNet encoder) [12]. Possible explanations:

- ViTs prioritize global context via self-attention; CNNs excel at local spatial detail via convolutions with inductive biases (translation equivariance, local connectivity)
- Patch-level SSL pretraining (256×256 or 512×512) doesn't capture fine-grained boundaries at pixel resolution

Potential solutions:

- Hybrid CNN-ViT architectures (e.g., ConvNeXt + Swin Transformer) combining local and global processing
- Dense prediction heads with multi-scale feature aggregation (FPN-style decoders)
- Pixel-level SSL pretext tasks (masked patch prediction at native resolution)

B. Loss of Global Spatial Structure

The dominant paradigm—pretrain on isolated patches, then aggregate via MIL—may discard crucial spatial information:

- Tumor-stroma interface topology (infiltrative vs. pushing borders)
- Spatial gradients in immune infiltration (peritumoral vs. intratumoral TILs)
- Long-range architectural patterns (cribriform growth, comedonecrosis)

Potential solutions:

- **WSI-level SSL**: Train directly on multi-resolution slide pyramids using hierarchical ViTs or spatial transformers that process entire slides end-to-end
- **Graph-based aggregation**: Explicitly model spatial relationships via GNNs with tissue-aware edge definitions
- **Sequence modeling**: Represent slides as ordered trajectories (e.g., raster scans, random walks) and apply sequential models (LSTMs, Transformers with positional encodings)

C. Data Bias and Generalization

Training datasets (TCGA, institutional cohorts) exhibit systematic biases:

Overrepresented:

- Common cancers: lung (NSCLC), breast (invasive ductal carcinoma), colorectal (adenocarcinoma)
- H&E FFPE slides (standard clinical workflow)
- Western patient demographics (North America, Western Europe)

Underrepresented:

- Rare cancers: mesothelioma, sarcomas, hematologic malignancies
- Special stains: PAS, Masson trichrome, Giemsa, immunofluorescence
- Cytology (Pap smears, fine-needle aspirates) and frozen sections
- Non-Western populations (diverse genetic backgrounds, epigenetic modifications)

This can degrade performance and raise fairness concerns when models are deployed to underrepresented domains. Multi-site, multi-country consortia are needed to curate diverse training datasets.

D. Interpretability and Clinical Trust

Foundation model embeddings live in high-dimensional latent spaces ($d \approx 768$–1024) that don't correspond to pathologist-interpretable features:

- Nuclear pleomorphism (Grade 1 vs. Grade 3 nuclei)
- Mitotic count (quantitative grading criterion)
- Cytoplasmic features (eosinophilia, vacuolization, mucin production)

Implications:

- Difficult for pathologists to validate model decisions or identify failure modes
- Reduced trust in clinical deployment, particularly for high-stakes decisions (cancer diagnosis, treatment selection)
- Limited ability to discover novel morphological biomarkers aligned with biological mechanisms

Potential solutions:

- **Concept bottleneck models**: Force intermediate representations to align with human-interpretable concepts (nuclear grade, gland formation, stromal proportion)
- **Prototype learning**: Identify representative exemplar patches for each class and compare test patches to prototypes
- **Feature attribution**: Saliency maps, integrated gradients, attention visualizations highlighting discriminative regions

E. Computational Accessibility

Training pathology foundation models requires:

- Large-scale WSI datasets (100k+ slides), often proprietary
- High-end GPU infrastructure (8× A100 or H100, multi-node clusters)
- Weeks of training time and significant energy consumption

This limits accessibility for smaller research groups, hospitals in low-resource settings, and developing countries. Open-source pretrained models (UNI, CTransPath) partially address this, but deployment at inference time can still be expensive for large patient cohorts.

Potential solutions:

- Model compression (quantization, pruning, knowledge distillation) to reduce inference cost
- Federated learning: Train models across institutions without sharing raw data, addressing privacy concerns while improving generalization
- Cloud-based inference platforms with pay-per-use pricing for low-volume users

F. Future Directions

1. WSI-level self-supervised learning. Train directly on full slides using hierarchical architectures (multi-resolution ViTs, spatial transformers) to avoid patch-then-aggregate paradigm and preserve global tissue organization.

2. Hybrid CNN-ViT architectures. Combine local convolutions (fine spatial detail) with global self-attention (contextual reasoning) to improve both classification and segmentation.

3. Causal modeling. Move beyond correlation (morphology predicts outcome) to causation (understanding *why* patterns drive outcome). Leverage interventional data (treatment response, tumor evolution) or causal discovery methods to identify mechanistic biomarkers.

4. Tighter multimodal integration. Joint training from scratch (rather than late fusion) of vision, genomics, and text encoders to learn shared representations optimized for multi-task prediction.

5. Standardized benchmarks. Community-wide evaluation protocols on diverse cohorts (rare cancers, external validation across multiple sites, few-shot settings) to enable fair comparisons and accelerate progress.

6. Federated learning. Collaborate across institutions without sharing raw data, addressing privacy regulations (GDPR, HIPAA) while improving generalization through diverse training cohorts.

IX. Conclusion

This review has synthesized the evolution of biomarker learning from whole slide images, tracing the field's progression from hand-crafted texture descriptors through CNN-based patch classification, attention-based Multiple Instance Learning, self-supervised representation learning, and recent pathology foundation models. Vision-centric models like UNI and Virchow demonstrate that large-scale histology pretraining yields robust, label-efficient, and cross-site-generalizable representations, with particularly strong performance on rare cancers and few-shot learning scenarios. Multimodal models like H-Optimus-0 show that integrating genomics and free-text reports can further improve molecular biomarker prediction, though practical deployment remains constrained by data availability and computational requirements.

Despite these advances, five fundamental challenges persist:

1) **Dense segmentation limitations**: ViT-based foundation models show only marginal gains (ΔDice $\approx$ 0.005) on pixel-level tasks, suggesting that global context prioritization may sacrifice fine-grained localization
2) **Loss of spatial structure**: The patch-then-aggregate paradigm may discard crucial tissue topology (tumor-stroma interfaces, immune gradients, architectural patterns)
3) **Data bias**: Training datasets overrepresent common cancers, H&E FFPE slides, and Western demographics, potentially degrading performance on rare entities and raising fairness concerns
4) **Interpretability gap**: High-dimensional embeddings don't map cleanly to pathologist-interpretable morphological features, hindering clinical validation and trust
5) **Computational accessibility**: Training foundation models requires 100k+ slides and multi-GPU clusters, limiting access for smaller groups and low-resource settings

Promising future directions include WSI-level self-supervised learning to preserve global tissue organization, hybrid CNN-ViT architectures balancing local and global processing, causal modeling to move beyond correlation toward mechanistic understanding, tighter multimodal integration via joint training, standardized community benchmarks for fair evaluation, and federated learning to enable privacy-preserving multi-institutional collaboration.

The field is advancing rapidly, with new foundation models and multimodal architectures emerging regularly. Translating these models from research benchmarks to clinical decision support systems will require addressing the persistent challenges outlined above, with particular emphasis on rigorous external validation, interpretability aligned with pathologists' reasoning, and equitable performance across diverse patient populations and tissue types. As computational pathology matures, the ultimate goal remains clear: to augment—not replace—pathologists' expertise, providing quantitative, reproducible, and scalable tools to improve diagnostic accuracy, accelerate discovery of novel biomarkers, and ultimately enhance patient outcomes.

References

[1] D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Med. Image Anal.*, vol. 58, 101544, 2019.

[2] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in *Proc. ISBI*, 2009, pp. 1107–1110.

[3] A. Vahadane et al., "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Trans. Med. Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[4] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. 3, no. 6, pp. 610–621, 1973.

[5] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. ICML*, 2018.

[6] M. Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nat. Biomed. Eng.*, vol. 5, pp. 555–570, 2021.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, 2020.

[9] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, 2021.

[10] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2023.

[11] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Med. Image Anal.*, vol. 81, 102559, 2022.

[12] R. J. Chen et al., "Towards a general-purpose foundation model for computational pathology," *Nat. Med.*, vol. 30, pp. 850–862, 2024.

[13] A. Saillard et al., "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature*, vol. 637, pp. panthology–and–rare–cancer–detection, 2024.

[14] J. N. Kather et al., "Pan-cancer image-based detection of clinically actionable genetic alterations," *Nat. Cancer*, vol. 1, pp. 789–799, 2020.

[15] N. Coudray et al., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, pp. 1559–1567, 2018.

[16] J. Saltz et al., "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Rep.*, vol. 23, no. 1, pp. 181–193, 2018.

[17] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. NeurIPS*, 2021.