Equivariant Neural Field based Whole-Slide representations for microsatellite instability prediction

Marga J. Don[*1] MARGAJDON@GMAIL.COM
David R Wessels[*1] D.R.WESSELS@UVA.NL
Ylva A Weeda[2,3] Y.A.WEEDA@AMSTERDAMUMC.NL
Hoel Kervadec[1,2] H.T.G.KERVADEC@UVA.NL
Sybren Meijer[2,3] S.L.MEIJER@AMSTERDAMUMC.NL
Erik J. Bekkers [1] E.J.BEKKERS@UVA.NL
[1] *University of Amsterdam, Amsterdam, The Netherlands*
[2] *Amsterdam UMC, Amsterdam, The Netherlands*
[3] *Cancer Center Amsterdam, Cancer Treatment and Quality of Life, Amsterdam, The Netherlands*

Editors: Under Review for MIDL 2025

Abstract

Determining microsatellite instability (MSI) from pathology whole-slide images (WSIs) is crucial for immunotherapy patient selection. While Deep Learning has succeeded in Colorectal Cancer (CRC), performance in Gastric Adenocarcinoma (GA) remains limited due to the subtle, less discriminative morphological features associated with MSI in gastric tissue. Existing Multiple Instance Learning (MIL) approaches typically rely on global patch descriptors, often failing to capture the fine-grained local geometric structures required for this task. In this paper, we propose a novel framework utilizing Equivariant Neural Fields (ENFs) for histopathology representation learning. Unlike conventional neural fields that compress signals into a single global latent vector, ENFs represent tissue patches as latent point clouds, explicitly grounding representations in local geometry and ensuring equivariance to rotation. We further propose a hierarchical pipeline where these patch-level point clouds are stitched to form a comprehensive WSI-level representation, which is processed by the Erwin architecture for slide-level prediction. We validate our method on the NCT-CRC-HE-100K dataset and a clinical GA cohort. Our experiments demonstrate that ENFs achieve superior reconstruction fidelity compared to non-geometric Neural Field baselines (MedFuncta) and produce highly informative representations that improve downstream MSI classification performance in challenging gastric adenocarcinoma cases.

Keywords: Pathology, Geometry, Equivariance, Neural Fields

1. Introduction

Pathology is the fundamental basis for cancer medicine, enabling diagnosis, biomarker assessment and effective treatment. This field is rapidly shifting from conventional microscopy to the digital assessment of whole-slide histopathology images (WSIs). Captured at extremely high resolutions digital WSI open new possibilities for the application of deep learning (DL) algorithms, particularly in light of the severe global shortage of trained pathologists. One area where DL algorithms have already proven valuable is in predicting biomarker status from WSIs. The application of Deep Learning (DL) to entire WSIs

[*] Contributed equally

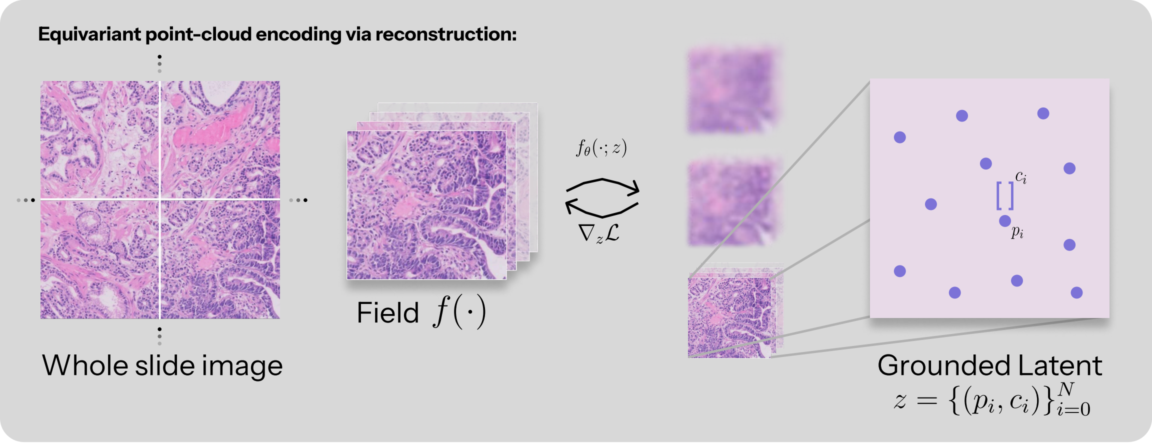

Figure 1: Equivariant Neural Fields (ENFs) ground neural representations in geometry using a latent point cloud. A latent set z consisting of tuples (p_i, c_i) of pose information p_i and context c_i is optimized to reconstruct to the image represented as (light) field $f(\cdot)$ as a function $f_\theta(\cdot; z)$ using gradient-descent. Due to their explicit positional grounding and locality, the latent retains important geometric features in the input field. The latent z can then be used in downstream tasks, on which will be elaborated in the paper.

is currently hampered by significant compute constraints due to their massive size. Consequently, standard practice involves splitting WSIs into numerous smaller patches to enable processing. To derive meaningful patch-level representations from this high volume of resulting data, methodologies like supervised learning are often employed (Wang et al., 2022). Previous approaches have already demonstrated that patch-level representations can be effectively leveraged for slide-level prediction tasks (Kather et al., 2019b; Echle et al., 2020; Wagner et al., 2023; Saillard et al., 2023). However, these patch-level representations are global descriptors and lack explicit information about the spatial organization of tissue structures within each patch. Though biomarkers are genetic features of tumors, their presence has clear and well-defined morphological effects on tumor tissue (Echle et al., 2020), making this a valuable feature for DL models to learn.

In this paper we utilize the recently proposed Equivariant Neural Field (ENF) framework (Wessels et al., 2024) for biomarker prediction. ENFs are Conditional Neural Fields that encode an image into a latent attributed point cloud—a set of position-feature pairs. This structure allows ENFs to ground latent features in specific spatial locations, thus capturing highly localized signal patterns. Importantly, this mapping is equivariant: geometric transformations of the input image are mirrored by corresponding transformations of the latent point cloud, ensuring the representation preserves the geometric structure of the underlying content. Since these latent representations are optimized to condition the shared backbone to reconstruct the original patch, they inherently capture patch-specific structural features. Furthermore, the patch-level representations can be stitched together to create a single point cloud representing the full WSI, allowing for models to consider the entire slide at once during inference, bypassing typical memory limitations that WSIs imply. This closely aligns with routine pathological assessment, during which pathologists consider spatial relationships across local regions and integrate these observations into a coherent slide-level evaluation.

Our main contributions are summarized as follows:

- We assess the feasibility of the approach by comparing the reconstruction performance of ENF to MedFuncta.
- We evaluate the medical applicability of this method for biomarker prediction in patients with gastric adenocarcinoma (GA). We demonstrate that the highly local and structural features captured by ENF, along with the ability to consider the full WSI simultaneously, allow for better prediction of Microsatellite Instability (MSI) than patch-based approaches.

2. Related works

We first introduce the medical context for the biomarker prediction task, explaining the biomarker we investigate and its clinical importance. Then, we introduce our main architecture, Equivariant Neural Fields, and describe its core underlying concepts. Finally, we introduce the architectures used in our downstream classification task.

2.1. Biomarkers in Immunotherapy and DL Prediction

Cancer can evade the immune system through mechanisms such as programmed ligand 1 protein (PD-L1) upregulation, which suppresses T-cell activation by binding to the PD-receptor. Checkpoint inhibitor therapy, a form of immunotherapy, aims to block the PD-1/PD-L1 interaction, restoring anti-tumor T-cell activity. The effectiveness of these therapies strongly depends on PD-L1 expression within the tumor tissue and the functionality of the DNA Mismatch Repair (MMR) system. When the MRR system is deficient, replication errors accumulate particularly in microsatellite regions, resulting in microsatellite instability (MSI). MSI/dMMR tumors typically have a high mutational burden and tend to respond well to checkpoint blockade, making MSI a key predictive biomarker recommended for testing in all gastric adenocarcinoma patients (Lordick et al., 2022). However, determining MSI status relies on immunohistochemistry analyses or molecular assays, which is labor-intensive and may delay therapeutic decision-making. Identifying MSI-associated morphological features directly on routine hematoxylin and eosin (H&E)-stained slides could therefore offer a faster, cost effective and more accessible alternative for early biomarker screening.

2.1.1. Previous Work on MSI prediction

Morphological features associated with MSI/dMMR tumors include an increased immune infiltrate (e.g. tumor infiltrating lymphocytes), mucinous or medullary differentiation, and poor tumor differentiation (Echle et al., 2020). These morphological features are visible on routine H&E slides and has motivated the development of deep learning (DL) models to predict MSI directly from histology. Early Convolutional Neural Network (CNN)-based approaches focused on patch-level classification (Kather et al., 2019b). Subsequent methods focused on aggregating patch-level predictions to a patient-level score, using simple averaging (Echle et al., 2020) or more sophisticated Multiple Instance Learning (MIL) (Saillard et al., 2023) and Transformer-based aggregators (Wagner et al., 2023). While these models have strong performance in colorectal cancer (CRC), their generalizability to GA is markedly lower (Kather et al., 2019b). This likely reflects that MSI-associated morphological patterns are more subtle in gastric tumors (Lee et al., 2023), making them harder to capture using models that rely on global tile-level features. By unifying multiple tile-level point clouds into a WSI point-cloud representation, ENF-based models preserve fine-grained spatial structure that traditional feature-aggregation approaches discard. Given that MSI-associated morphology in GA is subtle, we investigate whether ENFs can more effectively capture these nuanced patterns directly from routine H&E.

2.2. Equivariant Neural Fields

Neural Fields (NFs) (Tancik et al., 2020; Sitzmann et al., 2020) have emerged as a powerful paradigm for representing continuous signals. Unlike discrete arrays (e.g., pixel grids), NFs parameterize a signal as a function $f_{\boldsymbol{\theta}}$ mapping coordinates to values. This continuity offers resolution independence, but standard NFs require training a separate network for each data sample, which is computationally prohibitive for large datasets.

To scale this approach, **Conditional Neural Fields (CNFs)** (Park et al., 2019) separate sample-specific information from shared structure. They employ a shared backbone network conditioned on instance-specific latent vectors z. The **Functa** framework (Dupont et al., 2022) treats data *as* functions, using meta-learning to optimize these latent vectors such that they can reconstruct the original data. This reduces an entire dataset to a set of compact continuous function representations (”functaset”), enabling performant downstream tasks on the latent space. However, Functa typically uses a single global latent vector, which struggles to capture fine-grained, high-frequency details needed for complex visual tasks.

Addressing this, **Spatial Functa** (Bauer et al., 2023) reintroduced locality by assigning latent vectors to a grid of positions. While this significantly improved reconstruction and downstream performance, it reintroduced a discrete grid structure, sacrificing the resolution-agnostic benefits of the pure NF formulation.

Equivariant Neural Fields (ENFs) (Wessels et al., 2024; Knigge et al., 2024) resolve this dilemma by representing the latent space as a *point cloud* of feature-position pairs in a continuous domain. By treating latents as a set, ENFs maintain locality—grounding features in specific spatial regions—without relying on a fixed grid. Furthermore, ENFs explicitly enforce *steerability*, ensuring that geometric transformations (rotations, translations) of the input signal are exactly mirrored by transformations of the latent point cloud. This geometric grounding is particularly valuable for histopathology, where tissue architecture and cellular orientation are critical discriminative features often lost in rotation-invariant global descriptors.

3. Method

3.1. Datasets

We consider two datasets in this paper, highlighting the applicability of our method in both the patch-level and slide-level settings. The NCT-CRC-HE-100K dataset (Kather et al., 2019a) contains 100.000 non-overlapping patches extracted from 86 slides of human CRC tissue slides. Each patch belongs to one of nine tissue classes. Additionally, we consider the MSI-AUMC-SELECT-AI dataset, containing 12.175 patches extracted from 786 slides. These tissue sections consist of pretreatment biopsies and from gastric adenocarcinoma patients drawn from existing cohorts and treated at the Amsterdam Medical Centers. Patches were selected using tumor bulk segmentation, including a slide when over 50% of pixels contained tumor tissue. Contrary to NCT-CRC-HE-100K, the position of each patch within the original slide is available, thereby enabling both patch-level and slide-level operations.

3.2. Equivariant Neural Field Architecture

The core idea of the ENF architecture is to model the latent representation z as a point cloud $z = \{p_i, \mathbf{c}_i\}_{i=1}^{n_l}$, where $p_i = (\mathbf{p}_i, o_i) \in SE(2)$ specifies the position $\mathbf{p}_i \in \mathbb{R}^2$ and orientation $o_i \in S^1$ of a latent point, and $\mathbf{c}_i$ is an associated feature vector (see Figure 1). The steerability property ensures that for any group action $g \in SE(2)$:

$$\mathrm{ENF}_{\boldsymbol{\theta}}(g\mathbf{x}, gz) = \mathrm{ENF}_{\boldsymbol{\theta}}(\mathbf{x}, z) \tag{1}$$

This implies that operations depending on $\mathbf{x}$ and z jointly must be *bi-invariant*. We implement this by computing attributes $\mathbf{a}(\mathbf{x}_m, p_i)$ encoding relative pose. For $SE(2)$, this is $\mathbf{a}^{SE(2)}(\mathbf{x}_m, p_i) = \mathbf{R}_{o_i}^T(\mathbf{x}_m - \mathbf{p}_i)$. Locality is enforced via a Gaussian window $w_\sigma(\mathbf{x}_m, \mathbf{p}_i) = \exp -\frac{1}{2\sigma^2}\|\mathbf{x}_m - \mathbf{p}_i\|^2$ in the attention mechanism. These attributes are then embedded using a Gaussian Random Fourier Feature (RFF) embedding, $\phi(\mathbf{a}(\mathbf{x}_m, p_i))$, to alleviate spectral bias.

3.3. Latent Optimization and Whole-Slide Representation

We choose meta-learning (Finn et al., 2017) as our strategy to generate latents. We use $\mathbf{c}_i^0 = \mathbf{1} \in \mathbb{R}^{d_c}$ as the initial features for each latent. The latent positions are initialized on a grid, with added noise drawn from $\mathcal{N}(0, 0.01)$, to prevent overfitting to grid positions.

To create a slide-level latent representation, we stitch the patch-level point clouds into a single slide-level point cloud, resulting in $\{p_i, \mathbf{c}_i\}^{n_p \times n_l}$, where n_p is the number of patches in the slide and n_l the number of latent points in each patch. This slide-level point cloud can then be used for downstream classification using a point-cloud classification architecture.

3.4. Whole-Slide classification with ERWIN

The recently proposed Erwin (Zhdanov et al., 2025) architecture provides an efficient way to process large point clouds, combining the expressive power of attention with hierarchical processing. Other point-cloud methods which incorporate attention suffer from poor scaling, as attention is quadratic in the number of input nodes. Instead, Erwin computes self-attention over 'balls' of points, making attention linear in the number of balls. Originally, Erwin is designed to output a value for each input point, such as molecular dynamics or cosmological simulations. In order to achieve this, Erwin iteratively coarsens the point cloud to a bottleneck layer, and then iteratively refines the point cloud to the original number of points. In this work, we aim to predict slide-level attributes, using the embedding at the bottleneck layer. Therefore, we will not consider the refinement operation.

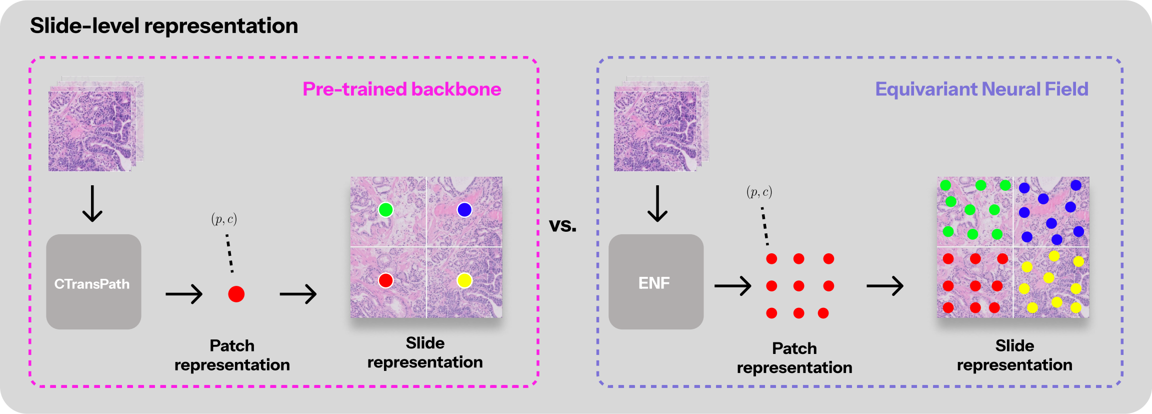

Figure 2: Instead of representing a whole-slide by different global tile-features extracted with a pre-trained backbone, we propose to utilise localised set-based features acquired by equivariant neural fields. By taking the union of multiple tile-based point-cloud representations, we acquire a point-cloud representation of the entire whole-slide.

Instead, we use the point clouds at the bottleneck layer as input to a classification head. Erwin requires the ball size $b = 2^n$, with n chosen as a hyperparameter. Since Erwin also ensures perfect binary trees at each layer, the point clouds at the bottleneck layer contain $b \cdot n_b$ points and feature vectors, with n_b the number of balls. Note that n_b is dependent on the size of the input point cloud, which, in the slide-level setting, is determined by the number of patches in the whole-slide.

Using the bottleneck features $\mathbf{c} \in \mathbb{R}^{(b \cdot n_b) \times d}$, with d the dimension of the feature vectors, we mean-pool within each ball to obtain $\mathbf{c} \in \mathbb{R}^{n_b \times d}$, and mean-pool again over the number of balls to obtain a single feature vector of size d for each point cloud. This feature vector is used as input to a 2-layer MLP to obtain a slide-level prediction. We present a schematic view of this pipeline in the right column of Figure 3.

4. Experiments

4.1. Datasets

We evaluate our method on two datasets, demonstrating its efficacy in both patch-level and slide-level tasks.

NCT-CRC-HE-100K (Kather et al., 2019a): This public dataset contains 100,000 non-overlapping image patches extracted from 86 hematoxylin and eosin (H&E) stained histological images of human colorectal cancer (CRC) and normal tissue. Each patch is labeled with one of nine tissue classes.

MSI-AUMC-SELECT-AI: This in-house dataset consists of 500 H&E stained WSIs from pretreatment biopsy (n = 246) and untreated resection specimens (n= 254). These slides were obtained from patients with GA who participated in various consortia at the

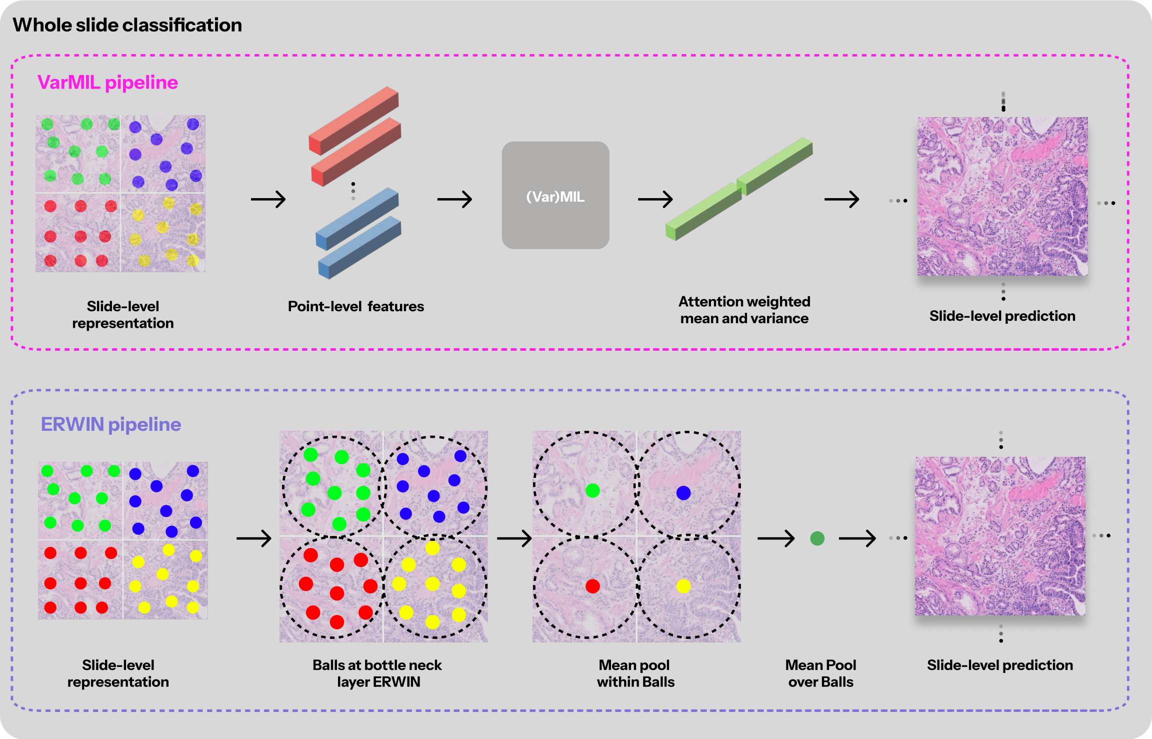
Whole slide classification
VarMIL pipeline
Slide-level representation
Point-level features
(Var)MIL
Attention weighted mean and variance
Slide-level prediction
ERWIN pipeline
Slide-level representation
Balls at bottle neck layer ERWIN
Mean pool within Balls
Mean Pool over Balls
Slide-level prediction

Figure 3: Overview of the slide-level classification pipelines. Left: The baseline approach extracts patch features using a pre-trained CTransPath encoder and aggregates them into a slide-level representation using VarMIL, which computes both weighted mean and variance. Right: Our proposed method utilizes ENF to generate a point cloud of latents for the entire slide. This large-scale point cloud is processed by the Erwin architecture, a hierarchical transformer that coarsens the representation to a bottleneck layer for efficient slide-level prediction.

Amsterdam University Medical Center between 1989 and 2023 and provided consent for reuse of their data. For 100 patients, both a biopsy and resection slide were available. All WSIs were digitized with a Philips IntelliSite Ultra-fast Scanner at a resolution of 0.25 µm/pixel and were manually inspected to ensure adequate image quality (e.g. high-resolution, sharp-focused). Each slide was accompanied by a ground-truth MSI label, with only 9.4% classified as MMR-deficient. Although this percentage is relative low, it aligns with expected clinical relevance. Given the class imbalance, particular care was taken to ensure that train/validation/test splits contained similar ratios of MSI versus non-MSI and biopsy versus resection specimens, both on slide- and patch-level. Slides were patchified into 256x256 tiles using the DLUP framework (Teuwen et al., 2024). All slides and patches originating from the same patient were assigned to the same data split.

4.2. Reconstruction task

We first validate the ENF framework on the patch-level reconstruction task, with the goal of learning a latent representation that captures the essential structural and morphological features of the tissue patches. To achieve this, we train the ENF model using Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017). In this framework, the model parameters $\boldsymbol{\theta}$ are optimized such that they can rapidly adapt to a new signal f_i within a few gradient descent steps (inner loop), producing a signal-specific latent representation z_i. We use 3 inner loop steps for both training and evaluation (see Appendix D for an ablation on inner loop optimization). The ENF model is configured with 64 latent points, each with a dimension of 64, and a backbone network with 128 hidden units.

We compare our method against MedFuncta, a state-of-the-art Neural Field approach for medical imaging. To ensure a fair comparison, we train the MedFuncta baseline on the same datasets using its official codebase, adapted for our specific patch sizes. We align the parameter budgets of both models to be comparable. The MedFuncta model is also trained using MAML with 3 inner loop steps, utilizing a Latent Modulated SIREN architecture with 10 layers and a latent modulation dimension of 512.

Our results, presented in Table 1, show that ENF achieves competitive reconstruction performance. This indicates that the latent point clouds successfully capture the detailed geometric structures present in the histopathology images, which is a prerequisite for effective downstream classification.

Table 1: Reconstruction results in PSNR (↑) and SSIM(↑)

	CRC Dataset		GA Dataset	
Model	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)
MedFuncta	16.17	0.879	NA	NA
ENF	23.41	0.729	22.44	0.661

4.3. Classification task

We verify our method on slide-level biomarker prediction on the GA dataset (Table 2). To evaluate the utility of the learned representations, we employ specific classification heads tailored to each method's output format. For the slide-level baseline, we extract patch features using a pre-trained CTransPath encoder, an architecture trained specifically for histopathology data, and aggregate them using VarMIL (Variance-based Multiple Instance Learning)(Schirris et al., 2022), as depicted in Figure 2 and Figure 3. VarMIL is an attention-based MIL pooling mechanism that considers both the weighted mean and variance of the patch features to form a slide-level representation. Finally, for our ENF model, which produces a point cloud of latents for each patch (or slide) (see Figure 2), we utilize the Erwin architecture (Zhdanov et al., 2025), as shown in Figure 3. Erwin is a hierarchical transformer designed for large-scale point clouds that efficiently aggregates local features through successive coarsening layers, producing a final embedding that is fed into a linear classifier.

We present the results of our experiments in Table 2. We observe that using ENF embeddings improves classification performance, most notably the F1 score. Furthermore, the combination of Erwin with ENF embeddings proves valuable, giving more robust performance and a much lower gap between performance on biopsy and resection tissue than Erwin with CTransPath embeddings. This implies that the highly local ENF embeddings capture valuable information for downstream classification. Notably, ENF achieves higher performance than CTransPath while using over 100x fewer parameters and over 1000x fewer GPU hours.

ENF-based prediction performed on-par with patch-based MSI-prediction on biopsy specimens, whereas ENF-based prediction yielded superior results for the resection specimens. This aligns with our hypothesis that ENFs are better suited to capture fine-grained spatial structure. Resection specimens preserve the broader architectural context of the tumor and surrounding tissue. In contrast, biopsies contain only a small, fragmented portion of the tumor, where the overall spatial orientation and layer structure are largely lost. As a result, the structural advantages of ENFs are less pronounced in biopsies but become highly beneficial in full-section tumor resections.

Table 2: Comparing feature extractor and downstream architecture baselines to ENF and Erwin. **Bold** reflects the best performance per metric, blue the second-best. For all metrics, higher is better.

Model	Feature Extractor		AUC %		F1 score %
		Overall	Biopsy	Resection	
VarMIL	CTransPath	50.2$_{\pm 0.6}$	55.5$_{\pm 0.8}$	42.0$_{\pm 0.6}$	5.8$_{\pm 8.3}$
	ENF	49.0$_{\pm 1.4}$	60.5$_{\pm 2.4}$	40.8$_{\pm 0.4}$	13.3$_{\pm 3.2}$
Erwin	CTransPath	60.4$_{\pm 9.9}$	**70.4$_{\pm 6.5}$**	57.7$_{\pm 24.2}$	12.6$_{\pm 2.2}$
	ENF	**63.9$_{\pm 4.6}$**	67.0$_{\pm 6.1}$	**60.0$_{\pm 2.5}$**	**21.5$_{\pm 4.1}$**

4.4. Exploring robustness in limited data settings

Compared to existing WSI datasets, such as those used to train CTransPath, SELECT contains only a fraction of the available slides. However, we are able to outperform CTransPath on the task of biomarker prediction on GC slides (Table 2. We aim to quantify this effect in this experiment, using a subset of the available slides in SELECT for ENF training. We included slides on a patient-level, keeping the balance of MSI/MSS consistent with the full dataset. ENF evaluation was performed on the unaltered test set, and classification training and evaluation was performed on the full dataset, to measure the effect of ENF as a feature extractor for unseen data.

Results are presented in Table 3, showing reconstruction performance decreasing slightly steadily but slightly as fewer patches are used for training. In classification performance, we mostly see decrease in resection performance, except in the case of using only 26.4% of patches, where we see an increase in resection AUC, but a decrease in biopsy AUC and overall F1 score. These results indicate that ENF generalizes well to unseen data, and highlights that a larger dataset could further boost performance of our method.

Table 3: Results of training ENF on a subset of the patches, selected at patient-level. Erwin is used as the downstream architecture. **Bold** reflects the best performance per metric, blue the second-best. For all metrics, higher is better.

% of patches used	PSNR	SSIM	AUC % Overall	AUC % Biopsy	AUC % Resection	F1 score %
100	**22.44**	**0.661**	65.7±2.8	68.0±3.8	63.6±2.4	22.0±2.9
76	22.26	0.658	**66.3±4.7**	**68.4±10.4**	62.8±3.6	**26.0±2.3**
53	22.07	0.629	64.8±8.0	66.0±9.0	60.9±2.9	24.9±5.3
26	21.60	0.577	61.6±2.3	61.7±2.4	**64.5±1.3**	20.6±0.2

5. Conclusion

In this paper, we introduced a novel framework for histopathology representation learning using Equivariant Neural Fields (ENFs). By representing tissue patches as latent point clouds, our method explicitly captures local geometric structures and ensures equivariance to rotations, a key property for analyzing histological tissues. We demonstrated that ENFs achieve competitive reconstruction fidelity compared to state-of-the-art Neural Field baselines while using significantly fewer resources.

For the clinical task of Microsatellite Instability (MSI) prediction in Gastric Adenocarcinoma (GA), our extensive experiments showed that ENF representations, when aggregated with the Erwin architecture, outperform standard global patch descriptors (CTransPath). Notably, our approach excelled in resection specimens, where it effectively leveraged the preserved spatial context to identify subtle morphological biomarkers. These results highlight the potential of geometrically grounded representations for analyzing complex tissue environments. Future work will focus on scaling to larger, multi-centric cohorts and extending the framework to other clinically relevant biomarkers.

Acknowledgments

This research was supported by a public grant from the 'Koningin Wilhelmina Fonds' (KWF150591) The funder had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript or in the decision to publish the results.

References

Matthias Bauer, Emilien Dupont, Andy Brock, Dan Rosenbaum, Jonathan Richard Schwarz, and Hyunjik Kim. Spatial functa: Scaling functa to imagenet classification and generation. *arXiv preprint arXiv:2302.03130*, 2023.

Emilien Dupont, Hyunjik Kim, SM Eslami, Danilo Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. *arXiv preprint arXiv:2201.12204*, 2022.

Amelie Echle, Heike Irmgard Grabsch, Philip Quirke, Piet A van den Brandt, Nicholas P West, Gordon G A Hutchins, Lara R Heij, Xiuxiang Tan, Susan D Richman, Jeremias Krause, Elizabeth Alwers, Josien Jenniskens, Kelly Offermans, Richard Gray, Hermann Brenner, Jenny Chang-Claude, Christian Trautwein, Alexander T Pearson, Peter Boor, Tom Luedde, Nadine Therese Gaisa, Michael Hoffmeister, and Jakob Nikolas Kather. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning, October 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.*, 16(1):e1002730, January 2019a.

Jakob Nikolas Kather, Alexander T. Pearson, Niels Halama, Dirk Jaeger, Jeremias Krause, Sven H. Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I. Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7): 1054–1056, July 2019b. ISSN 1078-8956. doi: 10.1038/s41591-019-0462-y.

David M Knigge, David R Wessels, Riccardo Valperga, Samuele Papa, Jan-Jakob Sonke, Efstratios Gavves, and Erik J Bekkers. Space-time continuous pde forecasting using equivariant neural fields. *Advances in Neural Information Processing Systems*, 37:76553–76577, 2024.

Sung Hak Lee, Yujin Lee, and Hyun-Jong Jang. Deep learning captures selective features for discrimination of microsatellite instability from pathologic tissue slides of gastric cancer. *International Journal of Cancer*, 152(2):298–307, 2023. doi: https://doi.org/10.1002/ijc.34251. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.34251.

F. Lordick, F. Carneiro, S. Cascinu, T. Fleitas, K. Haustermans, G. Piessen, A. Vogel, E.C. Smyth, and ESMO Guidelines Committee. Gastric cancer: Esmo clinical practice guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, 33(10):1005–1020, 2022. doi: 10.1016/j.annonc.2022.07.004.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

Charlie Saillard, Remy Dubois, Oussama Tchita, Nicolas Loiseau, Thierry Garcia, Aurelie Adriansen, Severine Carpentier, Joelle Reyre, Diana Enea, Katharina von Loga, Aurelie Kamoun, Stephane Rossat, Corentin Wiscart, Meriem Sefta, Michael Auffret, Lionel Guillou, Arnaud Fouillet, Jakob Nikolas Kather, and Magali Svrcek. Validation of msintuit as an ai-based pre-screening tool for msi detection from colorectal cancer histology slides, November 2023.

Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical image analysis*, 79:102464, 2022.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

J. Teuwen, L. Romor, A. Pai, Y. Schirris, and E. Marcus. Dlup: Deep learning utilities for pathology. https://github.com/NKI-AI/dlup, August 2024. Software.

Sophia J. Wagner, Daniel Reisenbüchler, Nicholas P. West, Jan Moritz Niehues, Gregory Patrick Veldhuizen, Philip Quirke, Heike I. Grabsch, Piet A. van den Brandt, Gordon G. A. Hutchins, Susan D. Richman, Tanwei Yuan, Rupert Langer, Josien Christina Anna Jenniskens, Kelly Offermans, Wolfram Mueller, Richard Gray, Stephen B. Gruber, Joel K. Greenson, Gad Rennert, Joseph D. Bonner, Daniel Schmolze, Jacqueline A. James, Maurice B. Loughrey, Manuel Salto-Tellez, Hermann Brenner, Michael Hoffmeister, Daniel Truhn, Julia A. Schnabel, Melanie Boxberg, Tingying Peng, and Jakob Nikolas Kather. Fully transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study, 2023. URL https://arxiv.org/abs/2301.09617.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

David R Wessels, David M Knigge, Samuele Papa, Riccardo Valperga, Sharvaree Vadgama, Efstratios Gavves, and Erik J Bekkers. Grounding continuous representations in geometry: Equivariant neural fields. *arXiv preprint arXiv:2406.05753*, 2024.

Maksim Zhdanov, Max Welling, and Jan-Willem van de Meent. Erwin: A tree-based hierarchical transformer for large-scale physical systems. *arXiv preprint arXiv:2502.17019*, 2025.

Appendix A. Learning the Latent Representation

CNFs and ENFs rely on a latent variable z that is optimized jointly with the shared decoder. In this work, we use *model-agnostic meta-learning* (MAML) to learn an initialization that can be adapted with a few inner-loop updates. Algorithm 1 provides the procedure used to obtain the latent representations in our reconstruction and downstream experiments.

Algorithm 1 MAML

```
Randomly initialize shared neural field $f_\theta$
for number of training iterations do
    Sample batch of signals $f$
    Sample random coordinates $x$
    Initialize latent vectors $z_f^{(0)}$ for all $f \in \mathcal{B}$
    for number of inner loop steps do
        Compute reconstruction loss $\mathcal{L}_{\text{MSE}}(f_\theta(x, z_f^{(t)}), f(x))$
        Update latent vectors:
            $$z_f^{(t+1)} \leftarrow z_f^{(t)} - \epsilon \nabla_{z_f^{(t)}} \mathcal{L}_{\text{MSE}} \quad \forall f \in \mathcal{B}$$
    end for
    Compute reconstruction loss across batch:
        $$\mathcal{L}_{\text{MSE}}^{\text{batch}} = \frac{1}{|\mathcal{B}|} \sum_{f \in \mathcal{B}} \|f_\theta(x, z_f^{(t+1)}) - f(x)\|^2$$
    Update neural field parameters:
        $$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \nabla_{\theta^{(t)}} \mathcal{L}_{\text{MSE}}^{\text{batch}}$$
end for
```

Appendix B. Latent evolution during meta-learning

In this section we show how the latent changes during each of the meta-learning optimization steps. Figure 4 shows the result of reconstructing the latent after each inner step. Note that most of the tissue structure is present after two optimization steps, which is reflected in the SSIM. The remaining steps seem to function mostly to optimize the color. Together, this indicates that increasing the number of inner steps would mostly increase reconstruction performance only, as the structural features necessary for classification are already captured by the latent in the first few steps.

Appendix C. Equivariance improves reconstruction quality

The bi-invariant attributes used in ENF are an essential aspect of the architecture. We investigate the effect of using an equivariance-breaking bi-invariant $\mathbf{a}^{\emptyset}$ and a rotational bi-

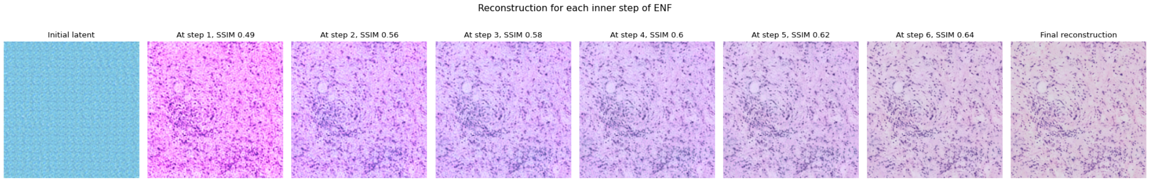

Figure 4: Example patch-reconstructions from latents at different steps of the meta-learning inner loop.

invariant $\mathbf{a}^{SE(2)}$, compared to $\mathbf{a}^{\mathbb{R}^2}$. In Table 4 we show results. We observe that equivariant processing ($\mathbf{a}^{\mathbb{R}^2}$ and $\mathbf{a}^{SE(2)}$) increases reconstruction performance slightly. Interestingly, though $\mathbf{a}^{SE(2)}$ latents are more expressive due to their added learned orientation, this does not translate into increased reconstruction or classification performance. We hypothesize that although the more expressive latent parametrisation, we needed a more restrictive bi-invariant to achieve $SE(2)$-equivariance. Moreover, the added expressiveness in the latent parametrisation is currently not carried through to the downstream architecture, which might explain the decreased classification performance for $\mathbf{a}^{SE(2)}$. Though the equivariance-breaking setting achieves a slight increase in resection AUC and F1 score, both metrics are less stable. Thus, $\mathbf{a}^{\mathbb{R}^2}$ remains the preferred bi-invariant.

Table 4: Results of using different bi-invariants during ENF training. Erwin used as the downstream architecture. **Bold** reflects the best performance per metric, blue the second-best. For all metrics, higher is better.

Bi-	PSNR	SSIM		AUC %		F1 score %
invariant			Overall	Biopsy	Resection	
$\mathbf{a}^{\emptyset}$	21.78	0.595	61.6±2.5	62.2±1.6	**63.8±5.1**	**24.1±5.1**
$\mathbf{a}^{\mathbb{R}^2}$	**22.44**	**0.661**	**65.7±2.8**	**68.0±3.8**	63.6±2.4	22.0±2.9
$\mathbf{a}^{SE(2)}$	22.36	0.649	62.0±4.7	64.1±5.6	61.1±4.6	18.5±2.9

Appendix D. Effect of Inner Loop Hyperparameters

Within the inner loop, latents are optimized for reconstruction and, ideally, such that they capture necessary features for downstream classification. As we will show in section D, the latent features are most influential in this process. In this experiment, we therefore investigate the optimization of latents within the inner loop, by increasing the number of inner steps and the learning rate for the features. Furthermore, we optionally work with learnable step-specific learning rates, as detailed in section **??**.

We present results in Table 5. Increasing the inner learning rate and number of steps clearly increases reconstruction performance. We note, however, that total runtime increases by 4 hours when performing 2 extra inner steps. Interestingly, the best model for overall classification has a clear bias towards biopsy slides, severely skewing the results and further indicating the challenge of classifying resection tissue specifically. We see that the model best suited for resection classification achieves the second-best PSNR and SSIM of this experiment, indicating that reconstruction performance can aid classification performance. Additionally, this model achieves a slight increase in the other classification metrics as well, leading us to use this configuration as a baseline in future experiments.

Table 5: Results of varying the number of inner step and inner feature learning rate, using Erwin as the downstream architecture. **Bold** reflects the best performance per metric, blue the second-best. For all metrics, higher is better.

Inner		Step-	PSNR	SSIM		AUC %		F1 score %
Steps	$\boldsymbol{\alpha}_c$	specific?			Overall	Biopsy	Resection	
3	80	55	21.83	0.605	$63.9_{\pm 4.6}$	$67.0_{\pm 6.1}$	$60.0_{\pm 2.5}$	$21.5_{\pm 4.1}$
		51	21.85	0.605	$63.9_{\pm 6.3}$	$65.2_{\pm 7.5}$	$62.2_{\pm 2.6}$	$19.7_{\pm 5.2}$
	200	51	22.03	0.624	$\mathbf{68.7_{\pm 2.8}}$	$\mathbf{73.0_{\pm 2.1}}$	$58.2_{\pm 1.4}$	$\mathbf{24.9_{\pm 3.8}}$
5	80	55	22.19	0.638	$60.7_{\pm 2.9}$	$63.5_{\pm 2.8}$	$58.5_{\pm 3.8}$	$19.6_{\pm 3.7}$
		51	22.19	0.640	$65.9_{\pm 0.4}$	$68.0_{\pm 1.3}$	$62.2_{\pm 3.7}$	$23.6_{\pm 1.6}$
	200	51	22.31	0.649	$62.8_{\pm 4.4}$	$64.9_{\pm 7.6}$	$62.4_{\pm 4.5}$	$21.2_{\pm 4.6}$
7	80	55	22.44	0.658	$65.7_{\pm 2.8}$	$68.0_{\pm 3.8}$	$\mathbf{63.6_{\pm 2.4}}$	$22.0_{\pm 2.9}$
		51	**22.45**	0.654	$63.9_{\pm 3.2}$	$69.8_{\pm 3.3}$	$58.3_{\pm 1.7}$	$23.0_{\pm 3.3}$
	200	51	22.42	**0.661**	$65.2_{\pm 3.3}$	$68.0_{\pm 4.7}$	$\mathbf{63.6_{\pm 4.4}}$	$18.4_{\pm 2.4}$

Appendix E. Effect of Number of Latents

The number of latent points modeled by the ENF is an important hyperparameter. As the latents are highly local, each latent essentially places a patch of pixels in its local area. Thus, increasing the number of latents would allow each latent to capture more detail, which we investigate in this experiment. Due to the latent initialization scheme chosen, we can only consider a square number of latents. For downstream experiments, the baseline Erwin architecture used in previous experiments fails when considering $n_l < 256$, since there are fewer points to create balls from. Thus, the architecture must be slightly changed to accommodate for this. We choose to keep 3 encoding layers and only vary the ball size in the last two layers, such that the smallest point cloud contains exactly 1 ball in the bottleneck layer. Additionally, we apply the Erwin configuration used with smallest n_l tested to all n_l tested, to ensure a fair comparison. Our results on the CRC dataset demonstrate that ENF representations yield competitive classification accuracy, affirming that the latent point clouds capture discriminative structural features. In the more challenging clinical setting of MSI prediction on the GA dataset, our method outperforms the baselines, highlighting the advantage of preserving local geometric information in the slide-level representation. Lastly, we examine the effect of the number of latents in each point cloud (see Appendix E), presenting the results in Figure 5. Increasing the number of latents leads to higher reconstruction performance, especially the SSIM, likely since each latent is now responsible for representing a smaller number of pixels. We show results in Table 6 and show the reconstructions of an arbitrary test image in Figure **??**. As expected, a higher number of latents vastly increases the reconstruction performance, especially the SSIM. This is

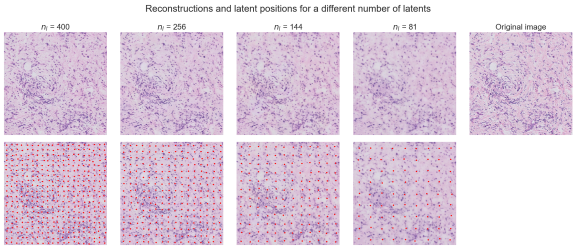
Reconstructions and latent positions for a different number of latents
n_l = 400
n_l = 256
n_l = 144
n_l = 81
Original image

Figure 5: Final reconstructions (upper row) and latent poses (bottom row, red dots) using different numbers of latents. The test image was chosen arbitrarily.

Table 6: Results for ENF runs using differing numbers of latents (n_l) per patch. In Erwin, we use ball sizes (128, 128) for the first 2 encoding layers and show the ball sizes for the last two encoding layers in this table. **Bold** reflects the best performance per metric, blue the second-best. For all metrics, higher is better.

n_l	PSNR	SSIM	Model	Ball Sizes	AUC % Overall	AUC % Biopsy	AUC % Resection	F1 score %
400	**23.41**	**0.729**	Erwin	128, 64	50.3$_{\pm 2.3}$	50.1$_{\pm 3.5}$	56.1$_{\pm 4.2}$	17.4$_{\pm 3.7}$
				64, 16	56.8$_{\pm 1.1}$	54.3$_{\pm 2.2}$	58.8$_{\pm 0.9}$	16.5$_{\pm 3.2}$
			VarMIL	-	53.3$_{\pm 2.6}$	64.4$_{\pm 2.6}$	46.8$_{\pm 5.2}$	19.2$_{\pm 6.5}$
256	22.44	0.661	Erwin	128, 64	65.7$_{\pm 2.8}$	68.0$_{\pm 3.8}$	63.6$_{\pm 2.4}$	**22.0$_{\pm 2.9}$**
				64, 16	**68.0$_{\pm 0.9}$**	**77.3$_{\pm 3.8}$**	56.0$_{\pm 6.7}$	21.1$_{\pm 0.1}$
			VarMIL	-	45.5$_{\pm 0.9}$	52.8$_{\pm 1.8}$	41.7$_{\pm 0.6}$	9.2$_{\pm 1.4}$
144	21.30	0.539	Erwin	128, 32	61.3$_{\pm 1.3}$	64.0$_{\pm 1.3}$	60.9$_{\pm 5.0}$	19.1$_{\pm 3.8}$
				64, 16	51.9$_{\pm 6.7}$	59.1$_{\pm 8.8}$	41.3$_{\pm 1.1}$	12.0$_{\pm 2.2}$
			VarMIL	-	48.8$_{\pm 0.2}$	57.9$_{\pm 2.1}$	49.3$_{\pm 0.6}$	12.7$_{\pm 1.3}$
81	20.49	0.431	Erwin	64, 16	56.0$_{\pm 5.6}$	51.1$_{\pm 8.1}$	**66.8$_{\pm 1.5}$**	17.2$_{\pm 4.8}$
			VarMIL	-	41.8$_{\pm 0.6}$	45.4$_{\pm 2.7}$	41.7$_{\pm 3.2}$	10.5$_{\pm 1.8}$

reflected in Figure **??**, where we see a higher number of latents indeed allows for much more detail to be captured. For classification, however, increasing the number of latents does not necessarily increase performance. When using VarMIL, best results are obtained using $n_l = 400$, but Erwin gives less clear results. Note that increasing the number of latents marginally increases ENF training time, ranging from 18 hours for $n_l = 81$ to 20 hours for $n_l = 400$.

As noted, the number of latents must be a square number, and $256 = 16^2 = 2^8$, meaning it is both a square and a power of 2. We hypothesize it is this property that leads to good synergy with the Erwin architecture, which assumes perfectly binary ball trees. However, more research is required to determine this with certainty.