



Advancing Atomic Force Microscopy Imaging with Machine Learning: Applications for Soft Matter

A dissertation submitted by
Mikhail Petrov

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Mechanical Engineering

Tufts University

August 2025

© 2025, Mikhail Petrov

Advisor: **Igor Sokolov**

Abstract

Advances in Atomic Force Microscopy (AFM) have enabled unprecedented nanoscale imaging and mechanical characterization of soft matter, but the complexity and volume of AFM data present significant challenges for analysis and interpretation. This thesis demonstrates how machine learning (ML) can enhance AFM as an experimental platform, with applications spanning biological and materials sciences.

The work begins by introducing the fundamentals of AFM, with an emphasis on sub-resonant tapping and Ringing Mode (RM), which enable simultaneous acquisition of multiple physical property maps and improved imaging throughput. The thesis then details the use of interpretable ML methods—specifically decision trees and random forests—for robust classification and feature ranking, allowing the extraction of physically meaningful surface parameters from high-dimensional AFM data. In biological applications, ML-driven analysis of AFM images is shown to accurately distinguish cancer cell types and aggressiveness, overcoming the limitations of traditional single-parameter approaches. By compressing AFM images into sets of surface parameters, the method achieves high classification accuracy even with limited data, and spatial heatmaps generated by the ML models enable localization of nanoscale features linked to disease phenotypes.

For materials science, the thesis introduces AFM-based mechano-spectroscopy (AFM-MS), which combines multi-channel AFM imaging with ML classification to identify and map constituent materials in complex samples at nanometer resolution. Two complementary approaches are presented: one leveraging databases of pure material properties, and another using known regions within a sample for training. The technique achieves lateral spectroscopic resolution down to 1.6 nm, surpassing conventional spectroscopic methods.

Finally, the thesis critically evaluates Bayesian deep transfer learning for small-data regimes, comparing standard transfer learning with advanced Bayesian approaches that use informative priors. It also examines innovative methods for hyperparameter selection, such as the data-emphasized Evidence Lower Bound (DE ELBo) approach, which learns regularization strength directly on the full training set without the need for com-

putationally expensive grid search. Systematic experiments across diverse datasets reveal that standard transfer learning remains a strong baseline, with the benefits of Bayesian methods being highly context-dependent. However, the DE ELBo method demonstrates comparable accuracy to existing approaches with significantly reduced compute time, particularly in limited data scenarios, offering a promising alternative for efficient fine-tuning.

Overall, this thesis establishes a comprehensive framework for integrating machine learning with AFM, enabling high-throughput, interpretable, and spatially resolved analysis of soft matter. The methodologies developed here advance both the automation and the scientific insight obtainable from AFM, with broad implications for diagnostics, materials design, and the future of intelligent microscopy.

Acknowledgments

I want to express my heartfelt thanks to my advisor, Igor Sokolov, who has been a constant source of support-whether I needed help with research questions or just some life advice. His encouragement and insightful feedback have made a huge difference in my journey.

I'm also grateful to my committee members, Shelly Peyton, Vivek Goyal, and Mark Kachanov, for taking on this important role in my academic path and for their thoughtful input along the way. Special thanks go to Mike Hughes and Ethan Harvey, who welcomed me into their Bayesian Deep Learning team and introduced me to this fascinating field.

Being part of such a supportive and close-knit group has been one of the highlights of my time here. I want to thank all my current and former lab-mates-Nishant Kumar, Haotian Jia, Yabo Cao, Paria Mir Hashemian, Dafu Wang, Mahshid Iraniparast, Renato Aguilera, Qing He, Nikita Kulachenkov, Rajendra Prasad, and Daniel Canena. Our group brings together people from all sorts of backgrounds, and I've learned something new from every one of you.

I also appreciate the help from the Tufts International Center, who were always quick to respond and incredibly supportive. Thanks to Mike Hughes and the Graduate Student Conference Reimbursement Fund for making it possible for me to attend NeurIPS.

Life abroad isn't always easy for international students, so I'm especially thankful to my American relatives-the Vlasov and Shpunt families-for their advice and kindness. My dogs Lundi, Shishka, and Hotch have provided their own special kind of emotional support (and a healthy disregard for my academic worries).

Most of all, I want to thank my wife, Zoia, for her patience, understanding, and love through the ups and downs of my PhD. And finally, my deepest gratitude goes to my mother, Liudmila Petrova, who has supported my dream of becoming a scientist from day one. Her wisdom and unconditional love have made all of this possible.

List of Figures

1.1	Schematics of AFM working principle.	3
1.2	A schematic of the signal exploited in ringing mode.	6
1.3	Schematics of a decision tree.	10
1.4	Confusion matrix.	18
2.1	The flowchart of the proposed ML method.	30
2.2	The regression classification of cells of two different aggressiveness.	31
2.3	Heatmaps of cells of two different aggressiveness levels.	32
2.4	An example of a combination of heatmaps with one of the AFM images.	34
2.5	Regions of low aggressiveness identified within a highly aggressive cell.	36
2.6	Examples of typical $10 \times 10 \mu\text{m}^2$ AFM images of precancerous and cancerous cells used in this study.	39
2.7	Schematics of machine learning (ML) analysis.	40
2.8	Results of the ML analysis distinguishing between precancerous and cancerous cells.	42
3.1	A schematic of the nano mechano-spectroscopy.	50
3.2	Details of the work of machine learning classification algorithms.	52
3.3	Experimental demonstration of the lateral spectral resolution of the described AFM-MS.	53
3.4	Comparison with confocal Raman microscopy/spectroscopy and mechano-spectroscopy: an example of a polystyrene sphere partially embedded in epoxy.	54
4.1	BTL: Error rate vs. target train set size on CIFAR-10	66
4.2	BTL: Empirical alignment of loss landscapes	67
4.3	DE ELBo: Model selection comparison	76
4.4	DE ELBo: Test-set accuracy on CIFAR-10 and Oxford-IIIT Pet	78

List of Tables

3.1	Comparison of two AFM-MS approaches	49
4.1	BTL: Transfer learning methods	64
4.2	BTL: CIFAR-10 heldout accuracy	68
4.3	BTL: Oxford Flowers heldout accuracy	68
4.4	BTL: Oxford-IIIT Pets heldout accuracy	68
4.5	BTL: FGVC-Aircraft heldout accuracy	68
4.6	BTL: HAM10000 AUROC on heldout data	68
4.7	DE ELBo: Possible priors	71
4.8	DE ELBo: Computational time comparison	77
4.9	DE ELBo: Accuracy on CIFAR-10	79
4.10	DE ELBo: Accuracy on Oxford-IIIT Pet	79
B.1	Distribution of labels in HAM10000 training set.	87

Contents

Abstract	i
Acknowledgments	iii
List of Figures	iv
List of Acronyms and Abbreviations	ix
1 Introduction	1
1.1 Atomic Force Microscopy	3
1.1.1 Imaging Modes	4
1.1.2 Measurement Capabilities	5
1.1.3 Ringing Mode	5
1.2 Decision trees-based machine learning algorithms	10
1.2.1 Decision trees	10
1.2.2 Random Forests	11
1.2.3 Feature Ranking	12
1.2.4 Classification probabilities	13
1.2.5 Monte-Carlo cross validation	14
1.2.6 Classification metrics	16
2 AFM-based Cancer Detection on Cellular Level	22
2.1 Introduction	22
2.1.1 Cell Sample Preparation: Fixation and Freeze-Drying	22
2.1.2 Nuances and limitations of cell measurements using AFM	23
2.1.3 Surface Parameters	24
2.2 Identification of Geometrical Features of Cell Surface Responsible for Can- cer Aggressiveness	27
2.2.1 Introduction	27

2.2.2	Methods	29
2.2.3	Results	33
2.2.4	Discussion	34
2.2.5	Conclusions	35
2.3	Distinguishing Precancerous and Cancerous Human Epithelial Cervical Cells	37
2.3.1	Introduction	37
2.3.2	Methods	38
2.3.3	Results	41
2.3.4	Discussion	42
2.3.5	Conclusions	43
3	Mechanical spectroscopy of materials using atomic force microscopy (AFM-MS)	47
3.1	Introduction	47
3.2	Methods	48
3.3	Results and Discussion	51
3.4	Conclusion	54
4	Bayesian deep transfer learning	59
4.1	Transfer Learning with Informative Priors	60
4.1.1	Introduction	60
4.1.2	Methodology	61
4.1.3	Results	65
4.1.4	Conclusion	69
4.2	Learning Hyperparameters via a Data-Emphasized Variational Objective	70
4.2.1	Introduction	70
4.2.2	Methodology	71
4.2.3	Results	77
4.2.4	Conclusion	79

A	Mathematical details	83
A.1	Second Derivative Test	83
A.2	Low-Rank Σ_p	84
A.2.1	Trace of the inverse	85
A.2.2	Squared Mahalanobis distance	85
A.2.3	Log determinant	85
B	Classification details	86
B.1	Dataset Details	86
B.2	Classifier Details	87
	Funding Disclosure	92
	Publication list	93
	Conference Presentations	94

List of Acronyms and Abbreviations

AFM-IR	Atomic Force Microscopy-Infrared
AFM-MS	Mechano-spectroscopic atomic force microscopy
AFM	Atomic Force Microscopy
AUC	Area Under the ROC Curve
BTL	Bayesian Transfer Learning
CNN	Convolutional Neural Network
DE ELBo	Data-emphasized Evidence Lower Bound
DL	Deep Learning
DT	Decision Trees
ELBo	Evidence Lower Bound
IID	Independent and Identically Distributed
IR	Infrared
KL	Kullback–Leibler
LR	Low Rank
MAP	Maximum A Posteriori
MCCV	Monte Carlo Cross-Validation
ML	Machine Learning
NLL	Negative Log-Likelihood
NN	Neural Network
RF	Random Forest
RM	Ringing Mode
ROC	Receiver Operating Characteristic curve
SGD	Stochastic Gradient Descent
SGHMC	Stochastic Gradient Hamiltonian Monte Carlo
TERS	Tip-enhanced Raman spectroscopy
VI	Variational Inference

Chapter 1

Introduction

One of the primary objectives of this thesis is to demonstrate how Machine Learning (ML) algorithms can enhance and reinforce Atomic Force Microscopy (AFM) as an experimental technique, with successful applications spanning diverse fields such as biological cells (Chapter 2) and polymers (Chapter 3). To provide a solid foundation, Section 1.1 introduces the fundamental principles of AFM, discusses its various operational modes, and outlines both its strengths and limitations. Special emphasis is placed on the subresonant tapping mode, complemented by the Ringing Mode (RM) modality, as these were primarily used to collect the data presented in this thesis. Section 1.2 presents an overview of Decision Tree (DT)-based methods, with a particular focus on Random Forest (RF), which served as the main algorithm for the projects described herein. Several key topics essential for interpreting the results of this work are discussed in detail:

- Section 1.2.3: Describes the feature ranking mechanism of RF algorithms
- Section 1.2.4: The interpretation of classification probabilities in DTs and RFs
- Section 1.2.5: Approaches to validating ML algorithms, including Monte Carlo validation
- Section 1.2.6: A summary of important classification metrics: accuracy, Area Under the Curve (AUC), sensitivity, specificity and confusion matrix

All experiments involving DTs and RFs were conducted using the Python package *scikit-learn* (**sklearn**) [1]. Throughout this text, I will reference this package and its relevant functions and parameters where appropriate.

1.1 Atomic Force Microscopy

Atomic Force Microscopy (AFM) is a powerful and versatile technique for imaging, measuring, and manipulating matter at the nanoscale [2]. Unlike traditional optical or electron microscopes, AFM does not rely on lenses or electron beams but instead uses a sharp mechanical probe to “feel” the surface of a sample, enabling three-dimensional imaging with atomic or near-atomic resolution.

The schematics of the AFM are shown in Fig. 1.1. At the heart of the AFM is a cantilever with a nanometer-scale sharp tip, typically made from silicon or silicon nitride, which acts as a sensitive force sensor. As the tip approaches and interacts with the sample surface, a variety of forces—including van der Waals, electrostatic, and mechanical contact forces—cause the cantilever to deflect.

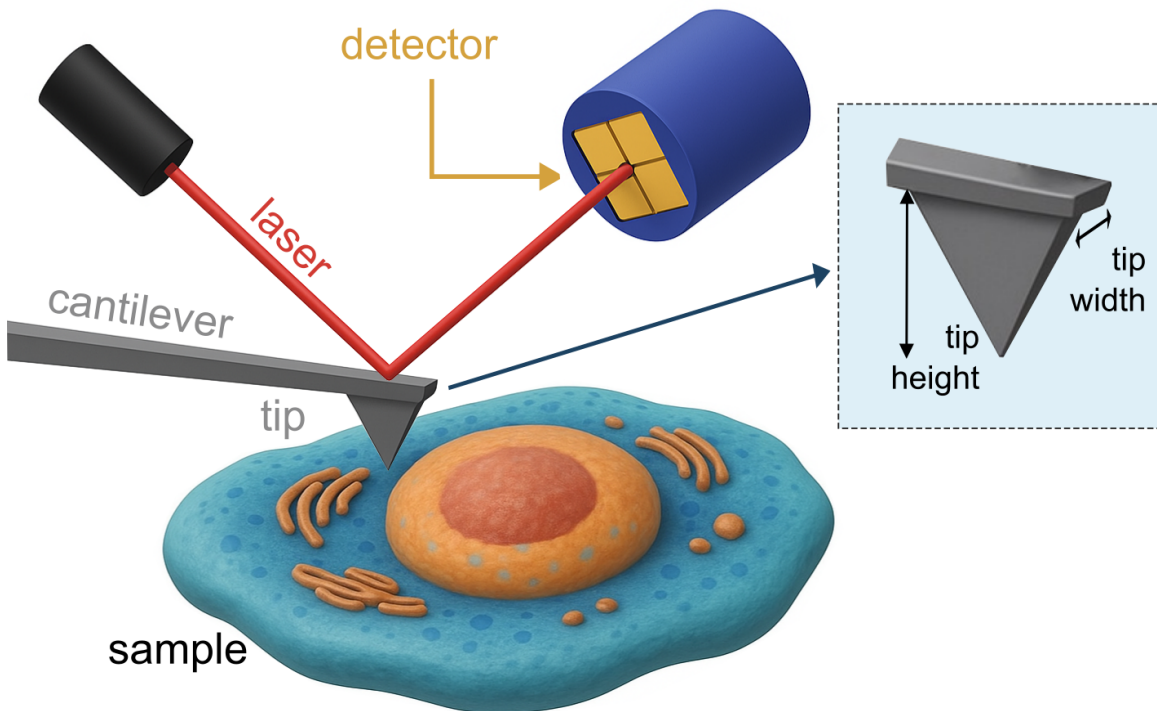


Figure 1.1: Schematics of AFM working principle.

These deflections are extremely small, often on the scale of nanometers or less, and are detected using a laser beam that is reflected off the top of the cantilever onto a position-sensitive photodetector. The change in position of the reflected laser beam corresponds to the cantilever’s deflection, which is directly related to the force between the tip and

the sample. The force can be quantified using Hooke's law ($F = -kd$), where k is the known spring constant of the cantilever and d is the measured deflection.

To create an image, the AFM scans the tip (or the sample) in a raster pattern over the area of interest. A feedback loop continuously adjusts the vertical position Z of the tip to maintain a constant interaction force (or another chosen parameter), ensuring that the tip follows the surface contours faithfully. The recorded vertical movements of the scanner as the tip tracks the surface are used to construct a high-resolution three-dimensional map of the sample's topography.

1.1.1 Imaging Modes

AFM can operate in several distinct modes, each characterized by the nature of the tip-sample interaction, which may involve constant contact, intermittent contact, or even no contact at all. Each mode is suited to specific types of samples and measurements:

- **Contact Mode:** The tip maintains continuous contact with the sample surface. The cantilever deflection is kept constant by the feedback loop, and the z-movements required to maintain this are recorded as the topography. While simple and effective, this mode can damage soft samples due to lateral forces.
- **Non-Contact Mode:** The cantilever oscillates just above the surface without actually touching it. Changes in oscillation amplitude or frequency due to long-range attractive forces are monitored to map the surface. This mode is gentler on delicate samples but can be more challenging in ambient conditions due to contamination layers.
- **Tapping Mode (Intermittent Contact):** The cantilever oscillates at or near its resonance frequency and periodically touches the surface. This reduces lateral forces and tip wear, making it ideal for imaging soft or loosely bound samples.
- **Subresonant Tapping Mode:** In this mode, the cantilever oscillates at frequencies well below its resonance frequency. The tip makes brief, controlled contacts

with the sample during each oscillation cycle, allowing precise control over the interaction force. Subresonant tapping modes—such as *PeakForceQNM*[™] [3]—are particularly advantageous for biological and soft matter samples, as they minimize damage, provide quantitative mechanical property mapping, and allow simultaneous acquisition of topography, adhesion, modulus, and energy dissipation data.

1.1.2 Measurement Capabilities

Beyond topographical imaging, AFM is capable of mapping a wide range of physical properties at the nanoscale. By modifying the tip or the operational mode, AFM can measure mechanical properties (e.g., stiffness, adhesion) [4], electrical properties (e.g., conductivity, surface potential) [5], magnetic properties [6], and even thermal characteristics [7]. This versatility is achieved by detecting how different forces between the tip and the sample affect the cantilever’s motion and by using specialized feedback and detection schemes.

AFM’s ability to operate in air, vacuum, or liquid environments makes it uniquely suited for biological, polymer, and material science applications. The technique provides true three-dimensional surface profiles with sub-nanometer vertical resolution and can image insulating, conducting, hard, or soft materials with minimal sample preparation. Notably, the spatial resolution of AFM is far superior to that of conventional optical microscopy, routinely achieving resolutions on the order of fractions of a nanometer—over a thousand times better than the optical diffraction limit [8].

1.1.3 Ringing Mode

Ring Mode (RM) builds upon the operational foundation of subresonant tapping modes. The innovation of RM lies in its use of the “ringing” signal—the residual oscillations of the cantilever that occur immediately after the tip detaches from the sample surface. Traditionally, this ringing signal (area highlighted by the ellipsoid in Fig. 1.2) has been treated as noise and filtered out in standard AFM modes. However, RM recognizes that these post-detachment oscillations encode rich information about the adhesive and

viscoelastic properties of the sample, particularly the molecular unbinding events that occur as the tip pulls away from the surface [9, 10].

Fig. 1.2 illustrates the basic working principle of the RM. Specifically, one full cycle of 1kHz vertical oscillation (subresonant tapping mode) of the Z scanner is presented. The lower half of the Fig. 1.2 shows specific points during the cantilever motion: (A) far from the sample surface, (B) touches the sample, (C) deforms the sample surface, (B*) touches the surface with zero force when retracting, (D) starts a fast detachment (pull-off) from the sample (the point defining adhesion force), (E) completely detaches from the sample (the point defining the restored adhesion force), (F) starts free oscillations above the sample surface

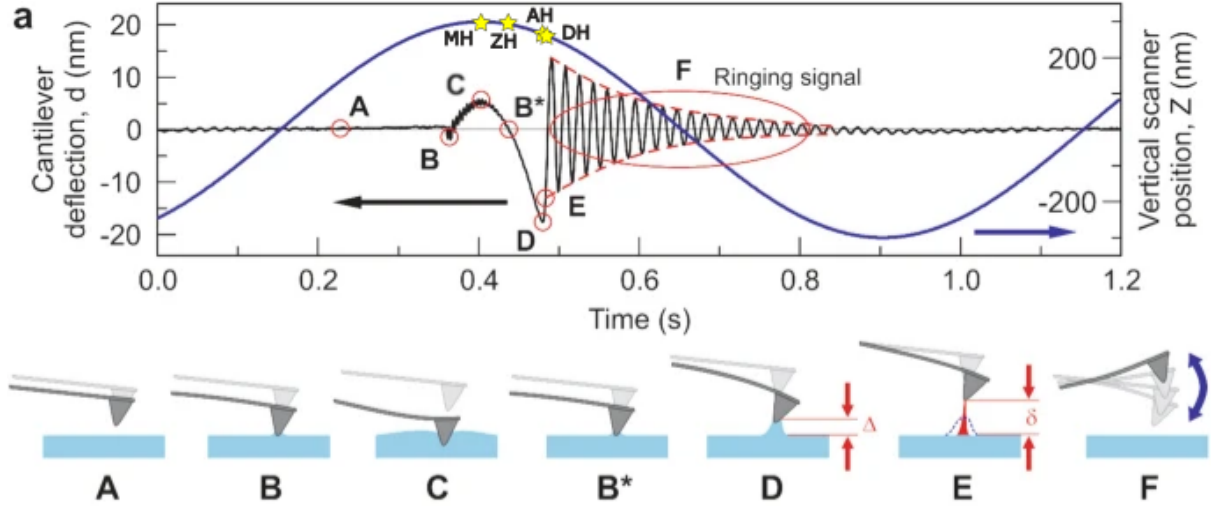


Figure 1.2: A schematic of the signal exploited in RM [9]. The cantilever deflection d and vertical position of the AFM scanner Z versus time are shown. The dash-envelope lines show the decrease of the oscillating amplitude of the AFM cantilever due to dissipation. Positive values of the cantilever deflection d are due to forces of repulsion, and vice versa, and negative values stand for the attraction between the AFM probe and sample surface. Scanner positions that are attributed to different height channels are indicated by star symbols.

By analyzing the ringing portion of the cantilever deflection signal, RM can extract a suite of new quantitative parameters in a single scan—up to twelve additional compositional imaging channels. These include:

- | | |
|----------------------------|-------------------------|
| • Traditional (max) Height | • Restored Adhesion |
| • Zero-force Height | • Viscoelastic Adhesion |
| • Adhesion Height | • Neck Height |
| • Disconnection Height | • Disconnection length |
| • Peak Force | • Disc Energy loss |
| • Adhesion | • Phase |

As shown in Fig. 1.2 with the star symbols, the four height channels are defined by the vertical scanner positions at distinct moments during the oscillation cycle. Max Height (MH) corresponds to point (C) on the force curve, where the pressing force reaches its maximum value, known as the peak force set point. Zero-force Height (ZH) aligns with point (B*) on the force curve, representing the moment when the force is zero during the cantilever's retraction. Adhesion Height (AH) is associated with point (D), which marks the maximal adhesion force, while Disconnection Height (DH) corresponds to point (E), the offset point of free oscillations of the cantilever.

Although these four channels are closely related in nature, each offers unique advantages. For instance, adhesion height is measured at the moment of maximum negative cantilever deflection, providing an image of the sample surface under pull-off stretching rather than compression. Through its inherently higher resolution, this channel enables the observation of surface corrugations—revealed by their stretching—that remain invisible in height channels recorded during compression.

Restored Adhesion measures the adhesion force between the AFM probe and the sample surface, accounting for residual molecular stretching after the pull-off point. Unlike regular adhesion, it captures the force after mechanical disconnection begins, reflecting the rupture of residual molecular bonds (Fig. 1.2 (E) points (D) to (E)). Immediately after the AFM tip detaches from the surface ($t = 0$) the cantilever behaves as a free,

lightly-damped harmonic oscillator. Its deflection is well-described by

$$d(t) = A_0 e^{-\gamma t} \sin(2\pi f_0 t + \varphi), \quad \gamma = \frac{\pi f_0}{Q}, \quad (1.1)$$

where A_0 is the undamped amplitude that existed at the instant of detachment, f_0 is the resonance frequency, Q the quality factor, and φ a phase that is zero at the first ringing maximum. All these parameters are determined using specific calibration techniques. Because each successive oscillation is attenuated by the exponential envelope $e^{-\gamma t}$, the peaks recorded at discrete times t_k are smaller than A_0 (area highlighted by the ellipsoid in Fig. 1.2).

To retrieve the original amplitude one inverts the decay by multiplying every measured deflection by the factor $e^{\gamma t_k}$. In addition, a geometric correction is required whenever the sampling point is not located exactly at a peak, because the sinusoid in (1.1) carries an extra factor $\sin(2\pi f_0 t_k + \varphi)$. Hence the reconstructed, or “inflated”, deflection that corresponds to the moment of detachment is obtained from

$$d_{\text{Ra}}^{(k)} = d(t_k) \exp\left(\frac{\pi f_0 t_k}{Q}\right) / \sin(2\pi f_0 t_k + \varphi). \quad (1.2)$$

All correctly reconstructed samples should equal the same value A_0 ; therefore the software averages them over several oscillations to mitigate electronic noise and timing errors,

$$\overline{d_{\text{Ra}}} = \frac{1}{N} \sum_{k=1}^N d_{\text{Ra}}^{(k)}. \quad (1.3)$$

Finally, multiplication by the calibrated spring constant k converts the reconstructed deflection into the restored adhesion force,

$$F_{\text{Ra}} = k \overline{d_{\text{Ra}}}. \quad (1.4)$$

Disconnection Length quantifies the height of molecular asperities or molecules stretched

by the AFM probe during final disconnection from the sample surface (Fig. 1.2 (E))

Pull-off Neck Height is the maximum height of the “neck” formed by the sample material stretched by the AFM probe during retraction (Fig. 1.2 (D)).

Disconnection Energy Loss measures the energy lost during the separation of the AFM probe from the sample, specifically due to the rupture of molecular bonds post-pull-off. It is defined as

$$E_{\text{dloss}} = \frac{k}{2}(d_{\text{a}}^2 - d_{\text{Ra}}^2), \quad (1.5)$$

where d_{a} and d_{Ra} are determined from the measured adhesion and restored adhesion forces $F_{\text{a/Ra}} = kd_{\text{a/Ra}}$.

Together, these channels provide a multidimensional view of the sample’s adhesive and viscoelastic landscape, significantly surpassing the information available from conventional force-distance curves or standard *PeakForce QNM*TM measurements [3]. For example, restored adhesion and detachment energy losses reveal subtle differences in molecular binding/unbinding dynamics, while pull-off neck height and detachment distance correlate with the stretching and relaxation of surface-bound molecules, such as extracellular matrix fibers or membrane tethers [9, 10]. Furthermore, the mode’s robust signal processing algorithms enable imaging speeds up to 20 times faster than traditional subresonant tapping, with reduced imaging artifacts and enhanced reproducibility.

1.2 Decision trees-based machine learning algorithms

1.2.1 Decision trees

Decision trees (DT) are non-parametric supervised learning algorithms used for classification and regression tasks [11, 12]. They operate by recursively partitioning the feature space into subsets based on decision rules, forming a hierarchical structure of nodes. A schematic view of a DT is shown in Fig. 1.3. At the root node, the entire dataset is evaluated, and splits are made at internal nodes using features that maximize class purity or minimize prediction error. Terminal leaf nodes represent final class labels or regression values.

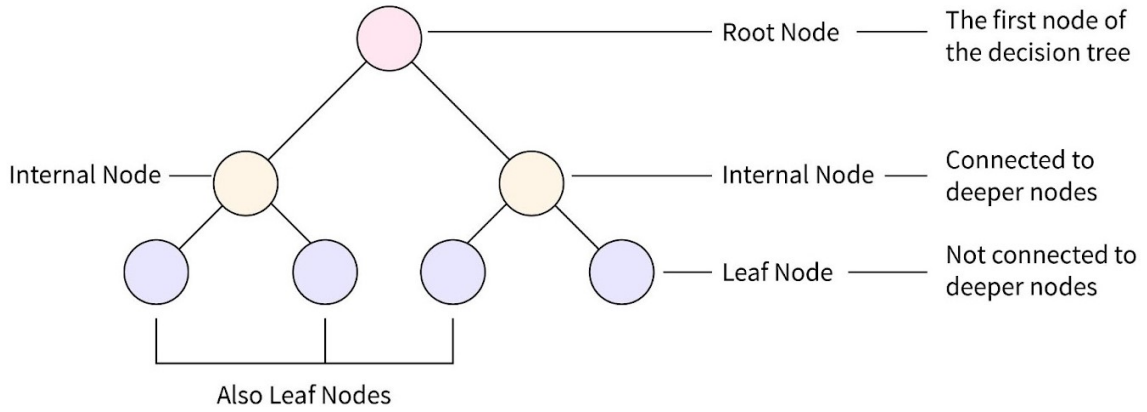


Figure 1.3: Schematics of a decision tree (DT) [13].

The learning process employs a greedy divide-and-conquer strategy. At each node, the algorithm evaluates all features and split points to identify the one that optimizes a chosen impurity metric. For classification, Gini impurity and entropy are commonly used. Gini impurity G quantifies the probability of misclassifying a randomly chosen element:

$$G = 1 - \sum_{i=1}^k p_i^2 \quad (1.6)$$

where p_i is the proportion of class i at the node, and k is the number of distinct classes.

Entropy H , derived from information theory, measures disorder:

$$H = - \sum_{i=1}^k p_i \log_2 p_i \quad (1.7)$$

Lower impurity values indicate purer nodes. For regression, variance reduction is typically used.

A critical challenge is balancing model complexity and generalization. Deep trees risk overfitting, capturing noise rather than patterns. Pruning addresses this by removing non-critical branches, either during training (pre-pruning) via constraints (e.g., max depth, minimum samples per leaf) or after training (post-pruning) by merging nodes that do not improve validation performance.

If the training data does not contain duplicate instances with different class labels (i.e., there are no conflicting samples with identical feature values but different classes), a DT can be grown without pruning to achieve complete purity. In this scenario, the tree can keep splitting on features until each leaf contains exactly one sample, ensuring that every leaf is perfectly pure.

1.2.2 Random Forests

Random forests (RF) are a widely used ensemble learning method that builds upon the strengths of decision trees (DTs) while addressing their tendency to overfit. Rather than relying on a single DT, a RF constructs a collection of individual trees, each trained on a different random subset of the data. This process, known as bootstrap aggregating or bagging, means each tree is exposed to a unique sample drawn with replacement (same data point can be chosen more than once) from the original dataset, introducing diversity among the trees and reducing overall variance [14].

During the construction of each tree, RF introduce further randomness by considering only a random subset of features at each split, rather than evaluating all possible features as in a standard DT. This random feature selection ensures that the trees in the ensemble are decorrelated, which enhances the model's ability to generalize and prevents any single

strong predictor from dominating the splits across all trees. As a result, the ensemble captures a broader range of data patterns and is less likely to overfit compared to a single deep DT.

For prediction, the RF aggregates the outputs of all its trees. In classification tasks, each tree casts a vote for a class label, and the majority vote determines the final prediction. For regression, the outputs of the trees are averaged to produce the final result. This aggregation leverages the “wisdom of crowds” effect, where the collective decision of many diverse models typically outperforms any individual model. RFs also provide measures of feature importance (Section 1.2.3), helping users understand which input variables contribute most to predictions.

It should be noted, however, that this improved accuracy and robustness come at the cost of increased computational complexity, especially as the number of trees grows, which can slow down prediction times in real-time applications.

1.2.3 Feature Ranking

The ability to estimate the importance, or ranking, of input features during the training process is a very important feature of RFs. This is especially valuable for cell classification (see Chapter 2), where we work with specific *surface parameters* (refer to Section 2.1.3) that correspond to distinct properties of the cell surface. By examining the most important features identified by the model, we can gain insights into which surface characteristics are most relevant for classification. Additionally, highly correlated features—which may be redundant for training—tend to share their importance scores. This property can be leveraged to identify and potentially eliminate such redundant features. The mechanism of feature ranking in RF operates as follows:

- **Impurity Reduction:** Each decision tree in the RF splits nodes based on features that best separate the data. The quality of a split is measured by the reduction in impurity (using metrics such as Gini impurity (Eq. 1.6) or mean squared error, depending on the task). When a feature is used to split a node, the reduction in impurity is attributed to that feature.

- **Summing Across Trees:** For each tree, the total decrease in impurity contributed by each feature is recorded. After all trees are built, these values are summed for each feature across the entire forest.
- **Averaging and Normalization:** The summed importance scores are then averaged and normalized so that the total importance across all features is 1. This produces a ranking in which higher scores indicate more influential features in the model’s predictions.

The calculated feature importances are accessible in most implementations (e.g., via scikit-learn attribute called `feature_importances_`). These importances are often visualized as bar charts, aiding in the selection of the most relevant features for the model.

1.2.4 Classification probabilities

Probability estimation is central to classification tasks. Classifiers use these probabilities to assign classes to new data. In neural networks, probabilities are typically calculated by applying a sigmoid (for binary classification) or softmax (for multi-class classification) function to the model’s outputs [15]. For decision trees (DTs), however, probability estimation is based on the class distribution within the leaf nodes, making the process slightly more nuanced.

Single Decision Tree: For a single DT, the probability of a class at a leaf node is determined by the proportion of training samples belonging to that class in the leaf.

$$P(\text{class} = c \mid \text{leaf}) = \frac{\text{Number of class } c \text{ samples in leaf}}{\text{Total samples in leaf}} \quad (1.8)$$

As discussed in Section 1.2.1, if a DT is grown without constraints (for example, with no `max_depth` limit), it can produce “pure” leaves, where all samples in a leaf belong to a single class. In such cases, the predicted probabilities at the leaves are either 0 or 1, which can lead to overconfident and poorly calibrated predictions. To mitigate this, trees are typically pruned or grown with constraints—such as setting `min_samples_leaf`

in scikit-learn—to ensure that leaves contain a minimum number of samples and may include multiple classes. This approach helps produce more reliable probability estimates by preventing leaves from becoming overly specific to the training data.

Random Forest: In a RF, final probability is computed by averaging probabilities from individual trees. Each decision tree in the forest outputs a probability vector (based on the class distribution in its leaf nodes). The final probability is the average of these vectors across all trees.

$$P(\text{class} = c \mid x) = \frac{1}{T} \sum_{t=1}^T P_t(\text{class} = c \mid x) \quad (1.9)$$

where T is the total number of trees, and P_t is the probability from tree t . This approach is used in scikit-learn’s `RandomForestClassifier` (via `predict_proba()`), especially when trees are grown to have impure leaves (e.g., `min_samples_leaf > 1`).

For single trees, it is best to rely on probability values only if trees are pruned (e.g., `max_depth=5`), as fully grown trees often produce overconfident 0/1 probabilities. For random forests (RF), probabilities are generally more reliable due to averaging across diverse trees.

1.2.5 Monte-Carlo cross validation

In the development of ML models, both validation and testing are essential for assessing model performance and ensuring trustworthy results. Validation involves evaluating the model on a subset of data not used during training, guiding model selection, hyperparameter tuning, and early detection of overfitting. This process helps practitioners refine their models and make informed decisions throughout development. In contrast, testing is performed only after all model choices have been finalized, using a separate test set that remains untouched during both training and validation. The test set serves as an objective benchmark, providing an unbiased estimate of how the model will perform on truly unseen data. By maintaining this clear separation between validation and testing, researchers can more accurately gauge the generalizability and real-world applicability

of their ML models. There are numerous methods for model validation, each with its own strengths and limitations. In this section, I will introduce the approach employed in the majority of my projects—Monte Carlo Cross-Validation (MCCV)—and discuss its relevance and application within this context.

MCCV, sometimes called shuffle-split cross-validation, is a resampling technique used to evaluate ML models. In MCCV, the dataset is randomly split into training and test sets multiple times, with the proportion of the split (such as 70% training and 30% testing) predefined by the user. Each iteration involves a new random split, and the model is trained and evaluated repeatedly, often dozens or hundreds of times. This approach provides flexibility, as it allows the user to control both the number of iterations and the size of the train-test split. The results from all iterations are averaged to obtain a robust estimate of model performance. Because each data point can appear in the test set multiple times across different splits, MCCV reduces dependency on any single partition and tends to produce lower variance in performance estimates.

Compared to other cross-validation strategies, MCCV offers several distinct characteristics. Its close relative, K-Fold cross-validation, divides the dataset into k equal-sized folds, using each fold as a test set exactly once while training on the remaining $k - 1$ folds. This ensures that every data point is used for both training and testing, but the number of unique splits is limited to k , and the results can be sensitive to how the data is partitioned, especially if the dataset is small or imbalanced. In contrast, MCCV can generate a much larger number of unique splits, which may provide a more comprehensive assessment of model stability, albeit at the cost of increased computational time.

Bootstrap validation is another alternative, where training sets are created by sampling with replacement from the original dataset, and the test set consists of the samples not selected in the bootstrap sample (the so-called out-of-bag samples). This method is particularly useful for small datasets and is closely related to bagging techniques in ensemble learning [16]. However, bootstrap samples often contain duplicate entries and may omit some data points entirely from the training set in each iteration, which can introduce optimistic bias in performance estimates.

While it is standard practice in ML to set aside a separate test set for final evaluation, the limited amount of available data in our case prevents us from doing so. Therefore, and given that Random Forest classifiers typically do not require extensive hyperparameter tuning, we rely on validation set results to report our model’s performance, acknowledging this as a limitation of our study.

1.2.6 Classification metrics

Evaluating the performance of ML classification models relies on robust measures that capture how well a model distinguishes between classes. This section provides a quick summary of the most important classification metrics [17–19].

Accuracy is the simplest and most intuitive metric, representing the proportion of correct predictions out of the total number of predictions made by the model. Mathematically, it is defined as the sum of true positives and true negatives divided by the total number of samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.10)$$

where TP , TN , FP , and FN stand for true positives, true negatives, false positives, and false negatives, respectively [17, 20].

- **True Positive (TP):** The number of instances where the model correctly predicted the positive class.
- **True Negative (TN):** The number of instances where the model correctly predicted the negative class.
- **False Positive (FP):** The number of instances where the model incorrectly predicted the positive class when it was actually negative (Type I error).
- **False Negative (FN):** The number of instances where the model incorrectly predicted the negative class when it was actually positive (Type II error).

In practice, accuracy is calculated after converting predicted probabilities into class labels using a chosen threshold, typically 0.5 for binary classification; thus, the reported accuracy can vary depending on this threshold. Accuracy is easy to interpret and is often the default metric for balanced datasets, as it provides a straightforward measure of overall correctness. However, in cases where the dataset is imbalanced—meaning one class is much more frequent than the other—accuracy can be misleading. For example, a model that always predicts the majority class could achieve high accuracy without actually learning to distinguish between classes [18, 19].

Sensitivity (or recall) quantifies the model’s ability to correctly identify positive cases, calculated as the proportion of true positives among all actual positives:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1.11)$$

Specificity measures the ability to correctly identify negative cases, representing the proportion of true negatives among all actual negatives:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (1.12)$$

To better understand the intuition of these metrics, we can consider a relevant to this thesis example of cancer detection:

Sensitivity is particularly valued, as failing to identify a cancer case (a false negative) can have severe consequences for patient outcomes. High sensitivity ensures that most patients with cancer are correctly flagged for further testing or treatment. For example, a sensitivity of 0.95 means 95% of cancerous cases are detected.

Specificity is also important, as a low specificity means more healthy patients may be incorrectly identified as having cancer, leading to unnecessary anxiety and follow-up procedures. A specificity of 0.90, for instance, indicates that 90% of non-cancerous cases are correctly ruled out.

Confusion matrix is another fundamental tool in machine learning used to evalu-

ate the performance of classification models. It is visually represented as a table that compares the actual class labels with the predicted class labels, providing a detailed breakdown of how well the model distinguishes between different classes and where it makes errors. The components of the confusion matrix are TP , TN , FP , and FN (Fig. 1.4).

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 1.4: Confusion matrix.

It is practical to note that when using `sklearn.metrics.confusion_matrix` for binary classification, the order of the values in the resulting matrix is different from the definition:

$$\begin{bmatrix} \text{True Negative (TN)} & \text{False Positive (FP)} \\ \text{False Negative (FN)} & \text{True Positive (TP)} \end{bmatrix}$$

Area Under the ROC Curve (AUC) provides a more nuanced evaluation of classification performance, particularly for imbalanced datasets. The Receiver Operating Characteristic curve (ROC) curve plots the true positive rate (TPR-also known as sensitivity or recall) against the false positive rate (FPR, which is equal to $1 - \text{specificity}$) across a range of threshold settings. The AUC summarizes this curve as a single value, representing the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one [18, 19]. An AUC of 1.0 indicates perfect discrimination, while an AUC of 0.5 suggests no better performance than random guessing.

Because AUC considers all possible classification thresholds, it provides a comprehensive view of model performance across different operating points and is particularly valuable when comparing models or dealing with class imbalance [17–19].

As an example, suppose we have two positive samples and two negative samples, with the following predicted scores(probabilities) from a classifier:

- Positive samples: $x_1 = 0.9$, $x_2 = 0.8$
- Negative samples: $y_1 = 0.7$, $y_2 = 0.4$

We consider all possible pairs (x_i, y_j) :

$$\begin{aligned}
 x_1 \text{ vs } y_1 : \quad & 0.9 > 0.7 \quad (\text{positive ranked higher}) \\
 x_1 \text{ vs } y_2 : \quad & 0.9 > 0.4 \quad (\text{positive ranked higher}) \\
 x_2 \text{ vs } y_1 : \quad & 0.8 > 0.7 \quad (\text{positive ranked higher}) \\
 x_2 \text{ vs } y_2 : \quad & 0.8 > 0.4 \quad (\text{positive ranked higher})
 \end{aligned}$$

All four pairs have the positive sample ranked higher. Thus, the empirical AUC is 1.0. If, for example, $x_2 = 0.6$, then $x_2 < y_1$ and only 3 out of 4 pairs would have the positive ranked higher, so AUC would be 0.75.

The ROC curve plots the true positive rate against the false positive rate at various thresholds. The AUC measures the area under this curve. Mathematically, it can be shown that this area is equal to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative one.

Bibliography

- [1] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [2] P. Eaton and P. West. *Atomic Force Microscopy*. Oxford: Oxford University Press, 2010.
- [3] Bruker Corporation. *PeakForce QNM*.
<https://www.bruker.com/en/products-and-solutions/microscopes/materials-afm/afm-modes/peakforce-qnm.html>. Accessed: 2025-04-29. 2022.
- [4] C. L. Essmann et al. “Mechanical properties measured by atomic force microscopy define health biomarkers in ageing *C. elegans*”. *Nature communications* 11 (2020), p. 1043.
- [5] X. Liang et al. “Simultaneous Visualization of Microscopic Conductivity and Deformation in Conductive Elastomers”. *ACS nano* 18 (2024), pp. 3438–3446.
- [6] V. G. Gisbert et al. “Quantitative mapping of magnetic properties at the nanoscale with bimodal AFM”. *Nanoscale* 13 (2021), pp. 2026–2033.
- [7] B. D. Iverson, J. E. Blendell, and S. V. Garimella. “Note: Thermal analog to atomic force microscopy force-displacement measurements for nanoscale interfacial contact resistance”. *Review of Scientific Instruments* 81 (2010).
- [8] *What is the resolution of an Atomic Force Microscope?* Accessed: 2025-04-30. 2025. URL: <https://www.afmworkshop.com/newsletter/what-is-the-resolution-of-an-atomic-force-microscope>.
- [9] M. E. Dokukin and I. Sokolov. “Nanoscale compositional mapping of cells, tissues, and polymers with ringing mode of atomic force microscopy”. *Scientific reports* 7 (2017), p. 11828.
- [10] I. Sokolov and M. E. Dokukin. “Imaging of soft and biological samples using AFM ringing mode”. *Nanoscale Imaging: Methods and Protocols* (2018), pp. 469–482.
- [11] J. R. Quinlan. “Simplifying decision trees”. *International Journal of Man-Machine Studies* 27 (1987), pp. 221–234.
- [12] J. Li et al. “Causal Decision Trees”. *IEEE Transactions on Knowledge and Data Engineering* 29 (2017), pp. 257–271.
- [13] Applied AI Course. *Decision Tree in Machine Learning*. Accessed: 2025-04-30. 2024. URL: <https://www.appliedaicourse.com/blog/decision-tree-in-machine-learning/>.
- [14] L. Breiman. “Random forests”. *Machine Learning* 45 (2001), pp. 5–32.
- [15] C. M. Bishop. “Neural Networks for Pattern Recognition”. *Oxford University Press* (1995).
- [16] Z.-H. Zhou. “Ensemble methods”. *Combining pattern classifiers*. Wiley, Hoboken (2014), pp. 186–229.
- [17] J. C. Obi. “A comparative study of several classification metrics and their performances on data”. *World Journal of Advanced Engineering Technology and Sciences* 8 (2023), pp. 308–314.

- [18] DataCamp. *AUC and the ROC Curve in Machine Learning*. 2024. URL: <https://www.datacamp.com/tutorial/auc>.
- [19] U. C. Science. *Lecture 9: Classification metrics*. https://ubc-cs.github.io/cpsc330-2023W1/lectures/09_classification-metrics.html. 2023.
- [20] A. Inc. *Classification metrics*. <https://apple.github.io/turicreate/docs/userguide/evaluation/classification.html>. 2024.

Chapter 2

AFM-based Cancer Detection on Cellular Level

2.1 Introduction

2.1.1 Cell Sample Preparation: Fixation and Freeze-Drying

For all the projects included in this thesis, prior to AFM measurements, all cells were fixed using Karnovsky solution and subsequently freeze-dried [1]. Working with live cells in AFM presents significant challenges due to their dynamic nature, softness, and sensitivity to environmental conditions. Live cells undergo constant mechanical and structural changes, such as membrane fluctuations and cytoskeletal reorganization, which introduce imaging artifacts and complicate the reproducibility of mechanical property measurements like stiffness and adhesion [2]. Additionally, maintaining physiological conditions (e.g., temperature, pH, osmotic balance) during imaging is technically demanding, and the risk of tip-induced damage to delicate cellular structures remains high [3].

In contrast, fixed cells provide stable imaging conditions by rigidifying cellular components through chemical crosslinking, enabling high-resolution surface topography mapping (e.g., microvilli, membrane pores) without motion artifacts [4]. Fixation simplifies mechanical analysis by eliminating dynamic cellular responses, allowing for direct compar-

ison of mechanical properties across samples [5]. While fixed cells lose some physiological relevance, they are compatible with longer imaging sessions, higher applied forces, and integration with complementary imaging techniques such as fluorescence or immunostaining [6].

Furthermore, cell lines are overwhelmingly preferred—and often essential—for most large-scale, reproducible, and regulated applications, primarily because of their ability to be preserved, expanded, and standardized [7]. Freeze-drying further enhances sample preparation by effectively removing moisture and preventing the accumulation of surface debris, both of which can otherwise compromise the quality of AFM imaging.

2.1.2 Nuances and limitations of cell measurements using AFM

AFM operating in subresonant tapping mode enables high-quality, reproducible cellular imaging through controlled force application and adaptive feedback mechanisms [8]. This technique achieves substantially higher lateral resolution than optical microscopy methods (0.1 – 1 nm vs. 200 – 400 nm) [9], permitting nanoscale imaging of cell surface physical properties [3]. The mode simultaneously acquires multiple data channels, such as Topographical mapping, Mechanical property quantification (e.g., stiffness, adhesion) and Energy dissipation analysis (see Section 1.1.1).

When integrated with Ringing Mode (RM) [6], the system expands from 3 additional measurable parameters in early implementations to 10 in advanced versions. This multimodal approach generates high-quality, multidimensional datasets by analyzing both tip-surface interactions and post-detachment cantilever oscillations (see Section 1.1.3).

Accurate AFM measurements rely on real-time feedback systems to maintain constant peak force, requiring multiple taps per pixel for optimal force regulation. This process inherently prolongs acquisition times due to the necessary stabilization period at each measurement point. Initial implementations using the torsional resonant mode *HarmoniX*TM [10] achieved precise cell surface characterization but required approximately 1 hour per cell scan¹ [11, 12]. Subsequent advancements transitioned to the subresonant

¹Typical cell scan was covering an area of $10 \times 10 \mu\text{m}^2$ with the resolution of 512×512 pixels.

tapping mode *PeakForceQNM*TM [13], which reduces imaging times to 20 minutes per cell. The improved speed-resolution balance in subresonant modes stems from synchronized probe trajectory tracking and reduced quality factor Q dependence [8].

Even with *PeakForceQNM*TM, the acquisition time remains a significant bottleneck, making it challenging to obtain very large datasets; as a result, the typical number of images is limited to a few hundred. Since our goal is to use ML to differentiate between different types of cells (e.g., malignant vs. non-malignant or malignant cells of varying aggressiveness) based on their AFM images, it is essential to establish an effective workflow that includes data preparation (see Section 2.1.3) and the construction of a ML classifier model. From a machine learning (ML) perspective—especially in the context of deep learning (DL)—each image is treated as a high-dimensional feature vector, where every pixel represents an individual feature. For instance, a standard 256×256 -pixel cell image generates 65,536 features per channel, creating a high-dimensional input space that rapidly becomes computationally intractable for state-of-the-art (SOTA) DL models—a phenomenon known as the “curse of dimensionality” [14]). These models are highly overparameterized, often containing tens of millions of learnable parameters [15], and require large datasets—typically from several thousand to millions of images—to achieve robust performance [16]. This challenge is further exacerbated as both the dimensionality of the images and the number of AFM channels increase². Given the limited size of our datasets, the direct application of DL methods is currently impractical. Chapter 4 describes DL models that are capable of operating in the scarce data regime.

2.1.3 Surface Parameters

Given the high-dimensional nature of atomic force microscopy (AFM) imaging data and the limited availability of samples, it is essential to adopt a dimensionality reduction strategy that balances data compression with interpretability. To address this challenge, we rely on surface metrology—the scientific discipline focused on quantifying surface topography using standardized parameters—as a foundation for our approach. These

²Three channels for a typical RGB image compared to ~ 10 channels for an AFM image.

industry-standard *surface parameters*, widely validated in engineering and materials science [17, 18], enable us to compress each high-resolution AFM image into a set of 35-40 numerical descriptors while preserving the physical meaning of the data. These parameters can be computed using commercial software such as MountainsSPIP[®] [19]; however, we developed our own Python-based library to provide greater flexibility and improved version control.

Surface topography describes the nanoscale deviations of a surface from an ideal flat plane, and these irregularities can have a profound impact on macroscopic properties such as friction, adhesion, and biocompatibility. Surface metrology provides a systematic framework for characterizing these deviations, typically dividing surface characteristics into amplitude parameters (e.g. Roughness and Mean Height), which measure vertical features; spatial parameters (e.g. Texture Directionality and Density of Summits), which capture horizontal patterns; functional parameters (e.g. Valley Fluid Retention Index), which describe aspects of surface texture related to performance or function (in terms of engineering application—how a surface will perform in terms of friction, abrasion, lubrication, and load support); and hybrid parameters (e.g. Mean Slope and Gradient), which integrate both vertical and horizontal aspects. While general-purpose dimensionality reduction techniques such as Principal Component Analysis (PCA) [20, 21] and DL-based autoencoders are available, they often obscure the physical interpretation of extracted features. In contrast, the use of *surface parameters* not only mitigates the curse of dimensionality inherent to high-resolution, low-sample-size datasets but also ensures that the resulting descriptors remain interpretable and relevant for both engineering and biological analyses. This domain-specific approach ultimately facilitates more meaningful statistical and machine learning analyses, as it preserves the link between quantitative image features and their underlying physical significance.

It has been shown that it was possible to segregate cancer and cells unambiguously using a single spatial surface parameter, fractal dimension, which was calculated using the maps of adhesion [11, 12]. However, the use of a single parameter had a serious limitation, and it was impossible to find such a precise separation between precancerous and

cancerous cells for which the differences are more nuanced. It is logical to expect that expanding the set of *surface parameters* should help with catching more subtle differences. In Section 2.3, we demonstrate that expanding the number of *surface parameters* from just one to six significantly enhances classification performance and enables reliable differentiation between precancerous and cancerous cervical cells. It is important to emphasize that each surface parameter possesses a well-defined physical meaning, which significantly enhances the interpretability of our ML model. This stands in contrast to many machine learning approaches that are often regarded as “black boxes”, where the relationship between input features and model decisions is not easily understood.

Due to the curse of dimensionality (briefly described in Section 2.1.2), image classification tasks that operate directly on raw pixel data are predominantly addressed using DL methods such as convolutional neural networks (CNN) [15]. These deep neural networks typically contain millions to billions of trainable parameters, rendering them computationally intensive to train and reliant on specialized hardware like GPUs for efficient optimization. Additionally, their high parameter count makes them prone to overfitting, particularly when training data is limited.

By contrast, our approach leverages *surface parameters* to compress high-dimensional imaging data into a low-dimensional feature space (reducing features from tens of thousands to just 30 descriptors). This compression enables the use of virtually any ML-based classification model, including logistic regression [22], decision tree-based algorithms (e.g., random forests [23]), and support vector machines [24]. Such simpler models are computationally efficient to train, require no specialized hardware, and are inherently less prone to overfitting due to their lower parameter complexity. Additionally, many classification algorithms rely on data standardization for efficient training. These algorithms are known as *scale-sensitive models*, which include, among others, CNN and support vector machines. Standardization is a data preprocessing technique that transforms the features of a dataset so that they have a mean of 0 and a standard deviation of 1. This is particularly important when different features have values that vary widely in scale. The process is also referred to as **Z-score normalization**, and is defined by the following

formula:

$$z = \frac{x - \mu}{\sigma} \quad (2.1)$$

where x is the original value, μ is the mean of the feature, and σ is the standard deviation. This procedure ensures that all features are placed on the same scale, allowing them to contribute equally during training. However, a drawback of this approach is that, for features with clear physical meaning—such as our *surface parameters*—standardization can reduce the interpretability of the results. For this reason, we can leverage our flexibility in algorithm selection and opt for *scale-invariant models*. One of the most popular scale-invariant approaches is decision tree (DT)-based models, such as Random Forest (RF), which are described in detail in Fig. 1.3. Lastly, the RF model provides feature importance scores (see Section 1.2.3), which allow us to identify the *surface parameters* most relevant for classification. This greatly enhances the interpretability of the results.

Even though the combination of *surface parameters* and the RF algorithm enjoys a relatively high level of interpretability, the data compression approach inevitably leads to some loss of information. Specifically, *surface parameters* characterize the entire measured surface, which may result in important local features being overlooked. In Section 2.2, we introduce a novel method to address this limitation by recovering local information and enabling correlations between cell-level classification and specific surface features.

2.2 Identification of Geometrical Features of Cell Surface Responsible for Cancer Aggressiveness

2.2.1 Introduction

Colorectal cancer is a leading cause of cancer-related mortality worldwide, with tumor aggressiveness being a key determinant of prognosis and therapeutic response. While conventional diagnostics focus on genetic and proteomic markers, recent advances highlight

the potential of physical and morphological cell properties as complementary biomarkers. Atomic force microscopy (AFM) offers nanoscale imaging of cell surfaces, providing high-resolution topographical and physical property maps [3, 12, 25].

Cell phenotypes are traditionally defined by their genetic signature and the expression of specific genes and proteins [26]. There are multiple attempts to characterize cells using physical methods, such as various optical spectroscopy [27–29], cell mechanics [2, 30, 31], and electron microscopy analysis [32, 33]. Recently, it has been demonstrated that atomic force microscopy can be used to image the cell surface with rather high precision and in a highly repeatable way [3, 12, 25, 34].

Our recent studies demonstrate that AFM, especially when combined with machine learning (ML), can distinguish cancer cell lines of varying aggressiveness with high accuracy [35]. However, as mentioned in Section 2.1, while AFM provides resolution far exceeding that of optical methods, the acquisition time becomes a limiting factor (approximately 20 minutes per single cell). This limitation results in inherently small datasets, which necessitate data space dimensionality reduction techniques. Our technique of choice is to compute a set of mathematical functions, called *surface parameters* (Section 2.1.3), that comprehensively describe the measured surface. *Surface parameters* retain enough information for accurate classification, but the downside is that all local information about specific regions of the cell surface becomes averaged out. This limitation prevents us from identifying which parts of the surface contribute to a particular classification outcome. AFM provides high-resolution images capable of resolving fine surface features, such as microvilli. Correlating these localized features with overall classification outcomes could offer valuable insights into how cancer manifests at the cellular level.

To address this, we introduce a machine learning framework capable of localizing geometrical regions on the cell surface critical for classifying colorectal epithelial cells as highly or lowly aggressive. By generating probability heatmaps across the cell surface, this method enables direct correlation of physical features with cancer aggressiveness, paving the way for future colocalization with molecular markers and deeper biological interpretation.

2.2.2 Methods

AFM Imaging of Cells in Ringing Mode

The AFM image dataset was assembled from two human colon cancer cell lines, HT29 and Csk, representing low and high aggressiveness, respectively [36]. Cells were imaged using the Ringing mode (see Fig. 1.2) of AFM, which allows simultaneous acquisition of multiple physical property maps, including height, adhesion, restored adhesion, and viscoelastic adhesion. RM utilizes the free oscillations of the AFM cantilever after detachment from the sample to extract information about viscoelasticity and adhesion energy [6].

For this study, three RM channels (RM adhesion, RM restored adhesion, and RM viscoelastic adhesion) and the height channel were analyzed. The adhesion channel was excluded due to artifacts. All images were acquired at 512×512 pixels over $10 \times 10 \mu\text{m}^2$ areas, and further processed as described below.

Machine Learning Workflow

The ML workflow used in this study is shown in Fig. 2.1. The core of our method is a random forest binary classifier [23], termed the backbone, which differentiates cells by cancer aggressiveness level. To train this classifier, we process AFM images by first splitting each 512×512 pixel image into four non-overlapping 256×256 pixel quadrants. For each quadrant, we compute ~ 40 *surface parameters* (see Section 2.1.3), then average these parameters across quadrants to yield one set per cell. This averaging mitigates the impact of imaging artifacts in AFM data. The surface parameters serve as features for binary classification. Once trained, the backbone classifier predicts whether a given cell belongs to the highly aggressive cancer class (Csk) or the low-aggressive class (Ht29). This process (Step 1 in Fig. 2.1) can use either a single AFM channel or multiple channels. For multi-channel training, surface parameters are computed per channel and concatenated into a single feature vector. As shown in Fig. 2.2(a,b), channel combinations provide a more comprehensive data description, improving class separation in classification probabilities.

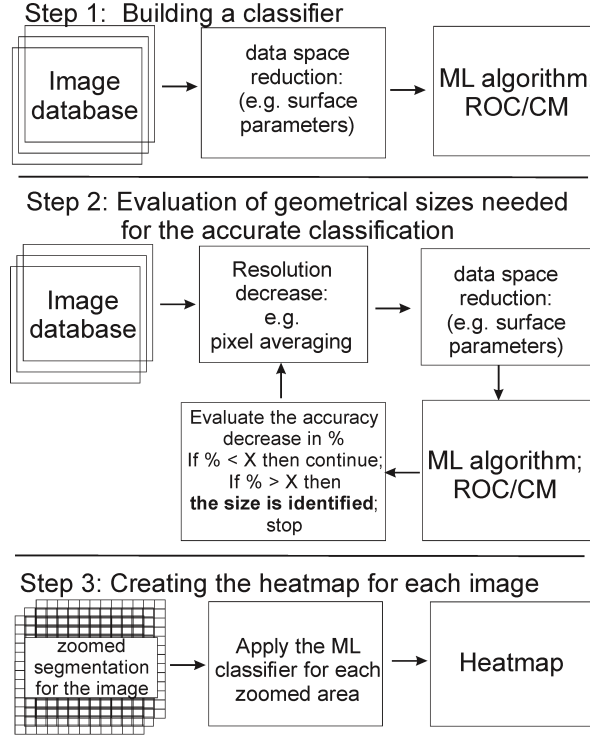


Figure 2.1: The flowchart of the proposed method. A suitable classifier is created in Step 1. The size of the zoomed area is defined in Step 2. The actual heatmaps are created in step 3 by applying the classifier algorithm developed in Step 1 to the zoomed-in areas of the size identified in step 2.

The classifier defined in Step I allows us to predict the cancer aggressiveness level for a given cell, but by converting our images into sets of *surface parameters*, we lose the ability to identify specific local regions on the cell surface that are responsible for a particular classification outcome. To restore some of this local information, we propose a method in which we extract small regions, or patches, from the original images and then use the backbone classifier to predict the cancer aggressiveness level for each patch. We choose the patches to be of fixed size and “slide” them over the original image with a step size of only one pixel. The backbone classifier’s prediction scores for all patches (one per patch) form an image that we refer to as a heatmap. The size of the patches is a crucial parameter in this method. The choice of patch size must be balanced: a large patch size will not produce sufficiently local heatmaps (resulting in low resolution), while patches that are too small may not retain enough information about the aggressiveness type. To determine the optimal patch size, we designed an experiment shown schematically as Step 2 in Fig. 2.1. The resolution of our AFM images is artificially reduced by applying an

2.2. IDENTIFICATION OF GEOMETRICAL FEATURES OF CELL SURFACE RESPONSIBLE FOR CANCER AGGRESSIVENESS

averaging kernel of varying size, after which *surface parameters* are recomputed for these blurred images and the backbone classifier is retrained on the new data. The effect of this, as can be seen in Fig. 2.2(c), is a drop in classification accuracy. The size of the blurring kernel can be associated with a certain length scale that is important for classification. In this way, we can define how small our patches can be while still retaining essential information. It can be seen that small blurring kernel sizes of 1–2 pixels (20–40nm) have a profound negative effect on classification accuracy. This length scale can be attributed to the distribution of receptors on the cell surface. However, we cannot choose our patches to be of such small size, simply because *surface parameters* lose their meaning at this scale. The next significant drop in accuracy, below 80%, occurs at a blurring kernel size of 32 pixels, which corresponds to around $0.6 \mu\text{m}$ and can be attributed to microvilli and microridges on the cell surface. We find that patches of this size work well, and so we choose to work with them.

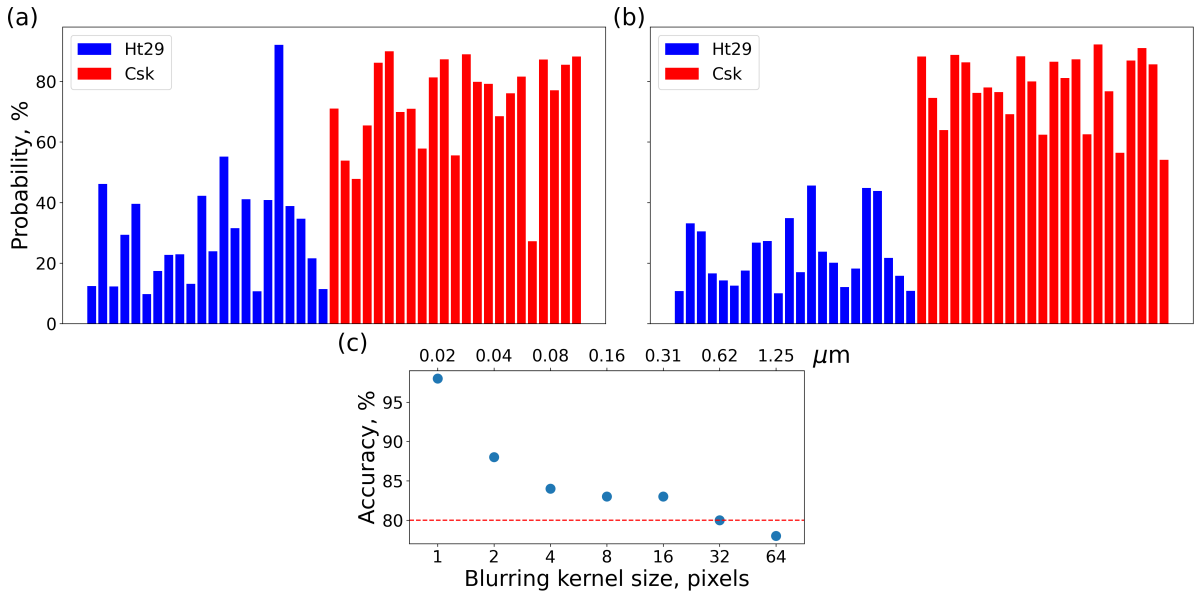


Figure 2.2: The regression classification of cells of two different aggressiveness when (a) using only one Ringing mode image (RM restored adhesion) or (b) using a combination of three Ringing mode images and the height of the image; (c) the decrease in the accuracy of classification (using the combination of channels) as a function of the pixel size of the averaging.

In Step 3 (Fig. Fig. 2.1), we generate heatmaps using the backbone classifier to predict outcomes for extracted patches. One important caveat arises: the backbone classifier was originally trained using *surface parameters* computed at $5 \times 5 \mu\text{m}^2$ (256×256 pixels),

while heatmap patches use parameters at $0.6 \times 0.6 \mu\text{m}^2$ (32×32 pixels). Since *surface parameters* are generally not scale-invariant, mixing these different length scales would introduce data incompatibility.

To resolve this, we retrain the backbone classifier. Each cell image is partitioned into non-overlapping 32×32 pixel subimages instead of quadrants. *Surface parameters* are computed per subimage and averaged per cell, yielding one parameter set per cell. Classification accuracy remains comparable to the original quadrant-based approach. This retrained classifier enables reliable heatmap generation from 32×32 pixel patches.

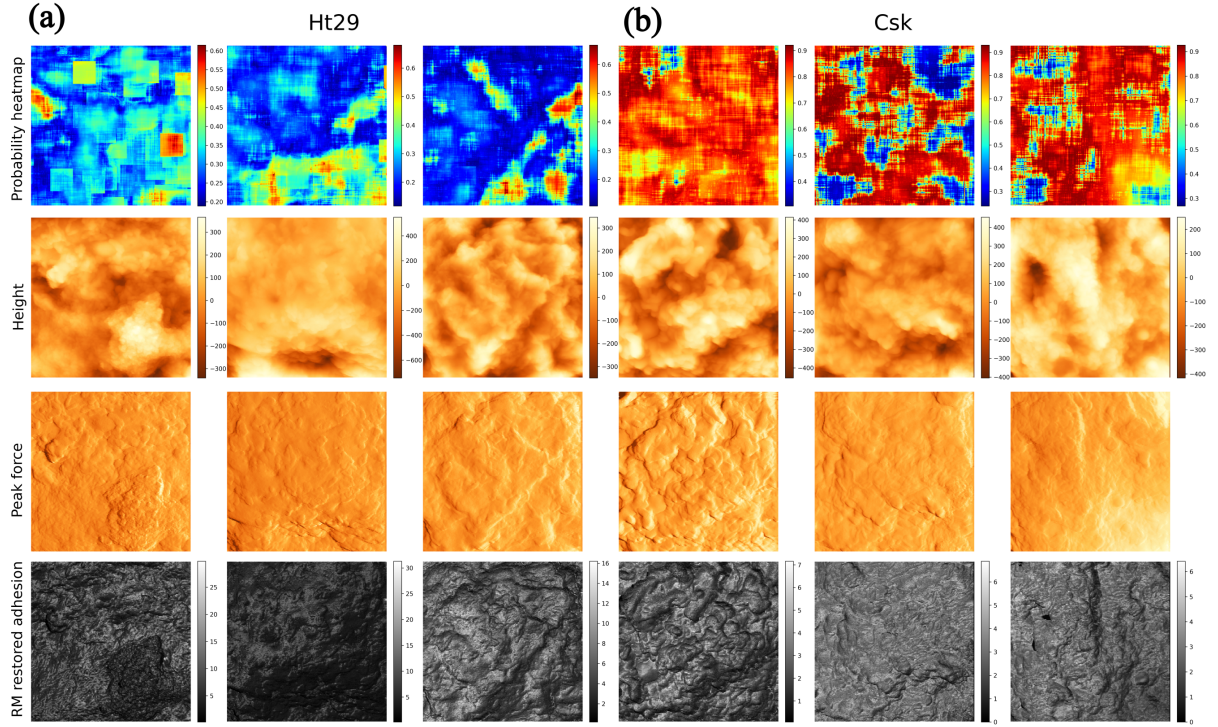


Figure 2.3: Heatmaps of cells of two different aggressiveness levels, low (a) and high (b), obtained when using one RM restored adhesion imaging channel. Three representative heatmaps for each class are shown. Each image represents $5 \times 5 \mu\text{m}^2$ recorded with 256×256 pixels (a quarter of $10 \times 10 \mu\text{m}^2$ recorded with 512×512 pixels). The top row shows the heat maps (probabilities for a 32×32 region around each pixel to belong to its class); the next row shows the height images of cells (the vertical scales are in nm); the third row shows the error of the feedback, PeakForce (in arbitrary units); and the bottom row shows RM-restored adhesion maps of the cells (the vertical scale is in nN).

2.2.3 Results

Classifier Performance and Feature Localization

Using only the RM restored adhesion channel, the classifier achieved high accuracy, with only four misclassified cells at a 50% probability threshold for high aggressiveness (Fig. 2.2 (a)). Combining three RM channels with the height channel improved classification to 94% accuracy at the single-cell level (Fig. 2.2 (b)).

Blurring experiments revealed that accuracy dropped significantly when features larger than 32×32 pixels ($\sim 0.5 \mu\text{m}$) were averaged, indicating that both nanoscale ($\sim 20\text{--}40 \text{ nm}$) and submicron ($\sim 0.5 \mu\text{m}$) features contribute substantially to classification performance (Fig. 2.2 (c)).

Heatmap Visualization and Correlation with Surface Features

Heatmaps generated by applying the ML classifier to overlapping zoomed-in regions across the cell surface identified spatially distinct patterns strongly correlated with aggressiveness (Fig. 2.3). Highly aggressive cells exhibited localized regions of consistently high classification probabilities, while low-aggressiveness cells showed inversely correlated spatial signatures. Cross-referencing these heatmaps with simultaneously acquired height and RM channel images revealed that areas of high or low aggressiveness probability frequently—though not exclusively—aligned with regions displaying characteristic microvilli or microridge topographies (Fig. 2.4). While the spatial correspondence between heatmap probabilities and surface features was significant, imperfect overlap in some regions suggests additional nanoscale factors may influence classification outcomes. This partial overlap indicates that, while localized surface architecture such as microvilli and microridges plays a significant role in determining aggressiveness, other nanoscale factors may also contribute to the classifier’s predictions.

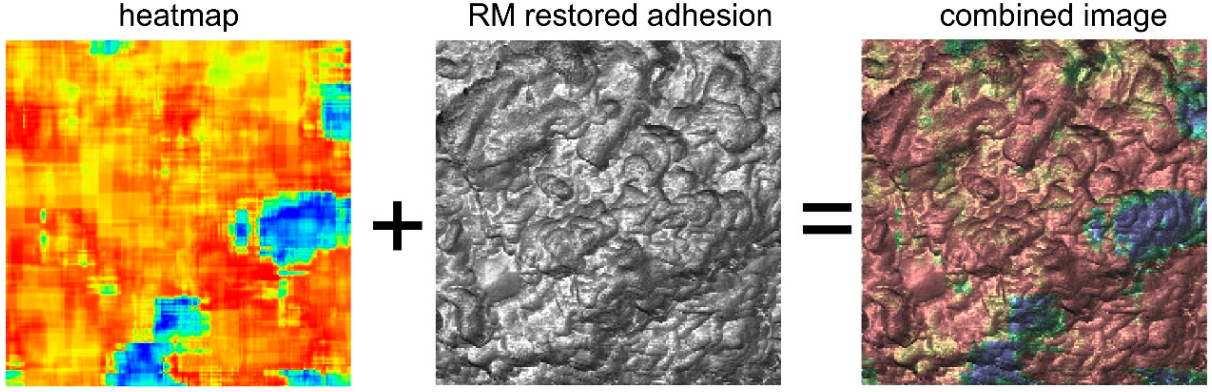


Figure 2.4: An example of a combination of heatmaps with one of the AFM images, RM restored adhesion. An example of a highly aggressive cell is shown.

2.2.4 Discussion

The presented method enables functional imaging of cell surfaces by integrating atomic AFM with ML to map physical properties linked to cellular aggressiveness. For instance, Fig. 2.4 demonstrates how heatmaps of RM-restored adhesion highlight regions of low aggressiveness, which could be synergized with fluorescence imaging to correlate physical features with molecular markers (e.g., cancer-associated receptors or proteins). However, the physical properties measured here—such as adhesion and topography—may reflect nonlocal molecular interactions or biochemical pathway combinations, suggesting complementary value to traditional fluorescence-based approaches [27–29]. As an example, regions of low aggressiveness on the surface of a highly aggressive cancer cell are shown in Fig. 2.5. Key surface features influencing classification accuracy include nanostructures ($\sim 20\text{--}40\text{ nm}$) and microscale geometries ($\sim 0.6\text{ }\mu\text{m}$). The former aligns with single-molecule dimensions (e.g., antibodies), while the latter may correspond to membrane protrusions like microvilli or actin-driven corrugations [12, 25, 31, 32]. Notably, averaging features beyond $\sim 0.6\text{ }\mu\text{m}$ reduces classification accuracy, implying that submicron structural details are critical discriminators. However, relying solely on geometry—such as microvilli density—proves insufficient, as highlighted by inconsistencies in low- and high-aggressiveness regions across cells. The method’s strength lies in combining multiple imaging channels (e.g., topography, adhesion, stiffness) rather than relying on a single parameter. For example, we observed improved accuracy when integrating four channels

compared to single-channel analysis (Fig. 2.3). This multi-parametric approach captures complex interactions between physical and biochemical properties, suggesting that future high-resolution imaging could refine feature localization.

A fundamental limitation arises in quantitatively validating patch-level predictions: the absence of ground-truth labels for individual subregions precludes direct evaluation of heatmap accuracy. As a result, performance metrics such as accuracy can only be reliably computed at the whole-cell level. Assessment of heatmaps therefore relies on correlative analysis, comparing spatial probability distributions with identifiable morphological features—for example, gradients in microvilli density—and, where possible, cross-referencing with orthogonal experimental techniques. This constraint highlights two important considerations.

The method primarily reveals physical surface properties rather than direct molecular identities. These physical characteristics emerge from complex molecular interactions and biochemical pathways. While colocalization with fluorescent labels is of interest, the approach likely captures information distinct from that provided by traditional biological assays. The nonlocal molecular origins of these surface properties suggest that our method is complementary to established biological characterization techniques and may offer synergistic insights.

Additionally, this approach shares conceptual parallels with Class Activation Mapping (CAM) in deep learning [37]. Like CAM-generated heatmaps, our method produces spatial probability distributions for which direct validation is not possible. In CAM, evaluation often involves masking high-probability regions and measuring the resulting decrease in whole-image classification confidence—a strategy that could, in principle, be adapted to our method to provide an indirect means of validating localized predictions, though this avenue remains to be explored.

2.2.5 Conclusions

This study demonstrates that AFM combined with machine learning provides a robust framework for differentiating between cancerous and healthy cells, as well as classifying

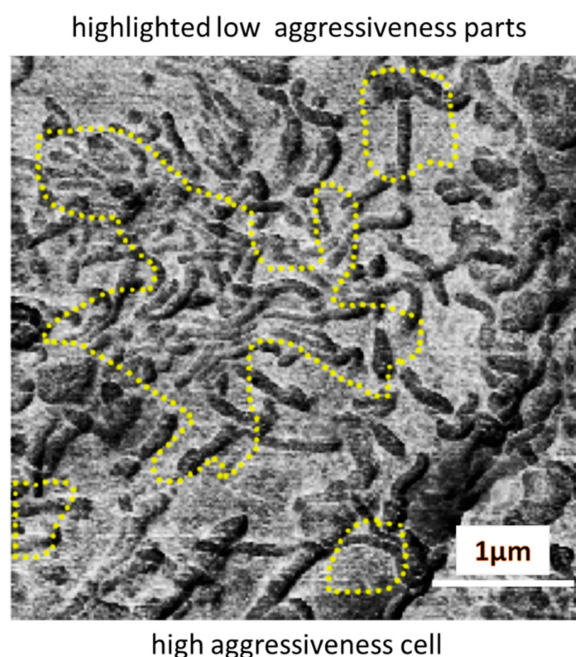


Figure 2.5: Regions of low aggressiveness identified within a highly aggressive cell. The regions are determined from and are superimposed on RM-restored adhesion channel image.

cancer cells by aggressiveness level. While achieving high classification accuracy, a significant limitation persists: the inherent challenge in identifying precise morphological or biophysical determinants driving these classifications. To address this interpretability gap, we developed a novel method that restores local spatial information lost during whole-cell feature extraction. This approach successfully correlates classification outcomes with specific cellular properties—such as microvilli density distributions—revealing that high-grade malignancy manifests as a local rather than global cellular characteristic. These findings establish a critical foundation for linking machine learning predictions with biologically meaningful surface properties at subcellular resolutions.

2.3 Distinguishing Precancerous and Cancerous Human Epithelial Cervical Cells

For this project, the data was collected prior to my PhD, and as a result, some of the novel approaches described in Section 2.1.2 Section 2.1.3 had not yet been implemented. In particular, AFM measurements were performed using the *HarmoniX*TM mode [10], rather than the faster and more advanced *PeakForceQNM*TM alternative [13]. Additionally, only six surface parameters were computed at the time, instead of the 35 described in Section 2.1.3. The results of this study are presented using various classification metrics, including accuracy, AUC, sensitivity, and specificity. Detailed descriptions of these metrics can be found in Section 1.2.6.

2.3.1 Introduction

Early detection is crucial for reducing the morbidity and mortality of cervical cancer, which remains the second most common cancer among women globally and a leading cause of cancer-related deaths (approximately 288,000 deaths annually and over 14,000 new cases diagnosed in the US in 2022 alone [38, 39]). Traditional screening with the Pap smear test has significantly lowered mortality rates by 70-80% [38, 39], but its sensitivity—especially for preinvasive (mean sensitivity $\sim 47\%$) and invasive lesions (16-82%)—remains suboptimal, leading to many equivocal results and unnecessary biopsies [39, 40].

DNA-HPV testing, now widely used due to the strong link between HPV and cervical cancer, also suffers from limited specificity and sensitivity, particularly in younger women where HPV infection is common but cancer is rare [40]. As a result, interest in improving Pap smear-based cytology persists. Despite being simple and minimally invasive, cytological tests have inherent limitations, including sampling and laboratory errors, and difficulties in reliably distinguishing between precancerous and cancerous cells using optical microscopy alone. These challenges highlight the need for more accurate

and reliable diagnostic methods.

To address these limitations, we employ AFM to analyze cervical epithelial cells, leveraging its superior resolution compared to optical microscopy [3, 12, 25]. Previous work demonstrated that fractal dimension analysis of AFM adhesion maps could distinguish normal from cancerous cervical epithelial cells, but with limited accuracy in separating precancerous from cancerous cells (AUC = 0.79, sensitivity = 58%) [25].

The current study introduces ML methods to analyze AFM adhesion maps, resulting in substantial improvements in diagnostic accuracy, sensitivity, and specificity. The new approach significantly reduces the false negative rate compared to conventional methods, addressing a critical gap in current clinical practice. The analysis is based on primary cell lines from different patients, underscoring the method’s potential for more effective and early cervical cancer detection [41].

2.3.2 Methods

Cell Lines and Sample Preparation

The study utilized twelve distinct human cervical epithelial cell lines, derived from tissues collected from the transformation zone of the cervix of six cancer patients and six healthy individuals, with tissue samples provided by the Cooperative Human Tissue Network. Cells were isolated through a two-stage enzymatic digestion process, first using enzyme to remove the epithelium and then trypsin to separate individual epithelial cells [42]. Precancerous (immortal) cell lines were generated by transfecting normal cells with HPV-16, which led to immortalization; non-immortalized cells ceased to proliferate after 60–150 doublings. All experiments were conducted on subconfluent cell cultures. The resulting cell lines included six precancerous and six cancerous variants, from which 108 precancerous and 64 cancerous cells were imaged using AFM. For imaging, cells were fixed with Karnovsky’s fixative, washed, and then dried using a freeze dryer before being stored in a desiccator. Imaging was performed at a relative humidity below 50% to ensure optimal conditions, and for each cell, a $10 \times 10 \mu\text{m}^2$ area was randomly selected for scanning to minimize operator bias. The dataset generated from these procedures is

publicly available online [43].

AFM Imaging

AFM imaging was conducted using a Bruker Dimension 3100 in *HarmoniX*TM [10], which captures adhesion force maps with nanoscale lateral resolution [41, 44]. Each cell was scanned over a $10 \times 10 \mu\text{m}^2$ area at 512×512 pixel resolution. Adhesion maps were divided into quadrants for analysis, excluding regions with artifacts [41]. Examples of AFM images are shown in Fig. 2.6.

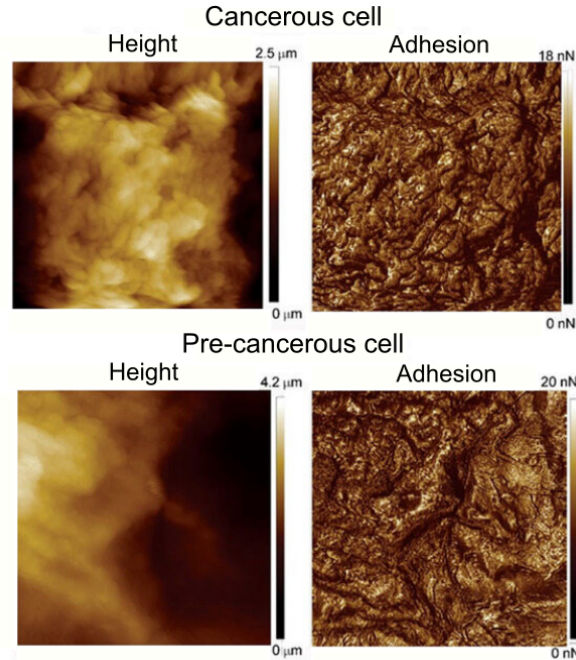


Figure 2.6: Examples of typical $10 \times 10 \mu\text{m}^2$ AFM images of precancerous and cancerous cells used in this study.

Surface Parameter Extraction

After AFM adhesion maps were collected, the first analytical step was to convert these high-resolution images into quantitative surface parameters for machine learning analysis (“Reduction of data” in Fig. 2.7). Specifically, six parameters were extracted from the adhesion maps: the developed interfacial area ratio (S_{dr}), which quantifies the percentage of additional surface area due to surface texture compared to an ideal plane; the root mean square gradient (S_{dq}), representing the RMS value of the surface slope and thus providing

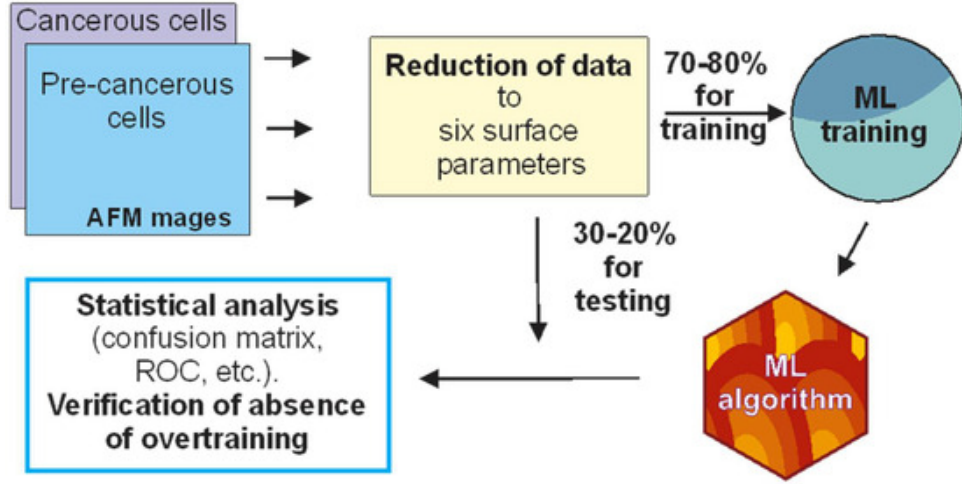


Figure 2.7: Schematics of machine learning (ML) analysis. Conversion of the AFM images into the *surface parameters*; splitting the database into the training and testing subsets; developing an ML algorithm using just the training subset; using the testing subset to perform the statistical analysis of the developed ML algorithm; and finally, cross-validation and the verification of the lack of overtraining of the developed approach.

information about the steepness and complexity of surface features; the reduced summit height (S_{pk}), which is derived from the bearing area ratio curve and reflects the height of the most prominent surface peaks; and three fractal dimension parameters—overall fractal dimension, (S_{fd_top}), and (S_{fd_bottom})—calculated through two-dimensional Fourier transform analysis of the adhesion maps [25, 41]. These fractal dimensions characterize the nanoscale roughness and hierarchical complexity of the cell surface, with S_{fd_top} and S_{fd_bottom} corresponding to different spatial frequency ranges in the AFM data.

Machine Learning Analysis

The schematics of the proposed ML approach is shown in Fig. 2.7. Random forest classifiers with 500 trees were trained and validated using Monte Carlo Cross-Validation (MCCV) with 500 random splits (see Section 1.2.5) to ensure robustness and avoid overfitting [23, 41]. Training/testing splits of 70/30% were used. As the next step, we sought to confirm that our classifier genuinely distinguishes between non-malignant and cancerous cells. To achieve this, we retrained the classifier on the same dataset, but with cell labels randomly reassigned. This approach tests whether the classifier is simply detecting random patterns in the data. Our results show that, when trained on randomly

labeled images, the classifier is unable to differentiate between cell types, thereby ruling out the possibility of overfitting to noise. Specifically, we performed 500 random label reassignments and, for each, calculated the classification metrics as the average across these random splits. By comparing the performance of 500 classifiers trained with true labels to those trained with randomized labels, we assessed the statistical significance of our results using an ANOVA test [45].

2.3.3 Results

The results demonstrate that the machine learning classifier achieved strong performance, with average values for true positives, true negatives, false positives, and false negatives of 0.91, 0.78, 0.22, and 0.09, respectively. The AUC value reached 0.93, while accuracy, sensitivity, and specificity were 83%, 92%, and 78% (Fig. 2.8 (a)). These values represent a significant improvement over previously reported results [25], especially in sensitivity, which increased from 58% to 92%. This improvement is particularly important in the clinical context, as current noninvasive screening tests for cervical cancer, such as the Pap smear and HPV tests, suffer from low sensitivity, resulting in many unnecessary invasive biopsies. The high sensitivity achieved by the classifier suggests that it could greatly reduce the number of missed cancer cases and unnecessary procedures if translated into clinical practice. The AUC is highlighted as the most reliable measure of improvement, as it is independent of the probability threshold used for classification, which can otherwise shift the balance between sensitivity and specificity. The false-negative rate for the current method is about 7%, a substantial improvement over the 27% rate reported for modern invasive colposcopy tests. However, these results are based on cell lines, so further validation with patient-derived samples is necessary before clinical application.

To confirm that the classifier was not overfitting, a control experiment was performed in which class labels were randomized; under these conditions, the classifier's performance dropped to chance levels, with an AUC of 0.5 and accuracy, sensitivity, and specificity around 50% (Fig. 2.8 (b)). This control, repeated 500 times, served as a baseline for statistical comparison. Using a one-way ANOVA test [45], the difference between the

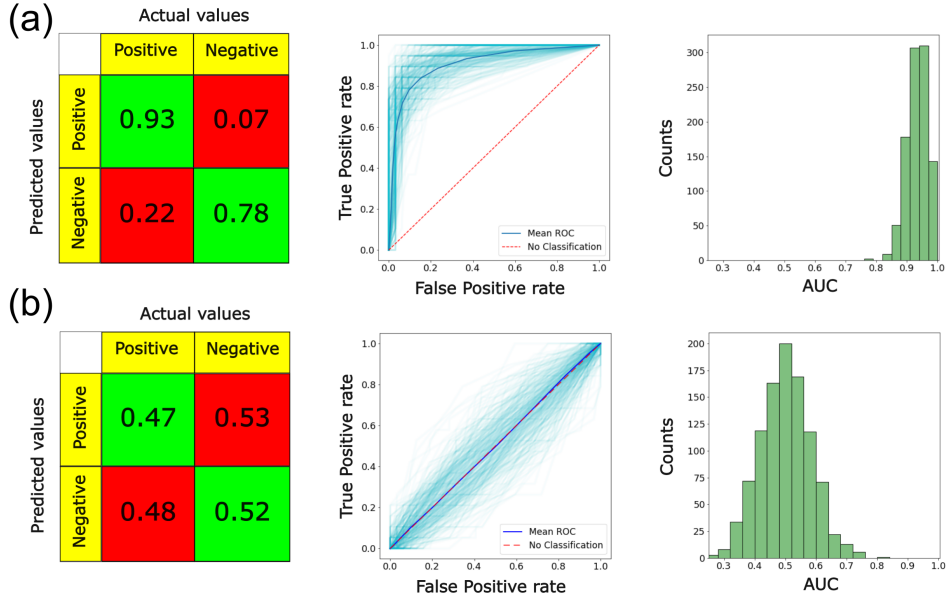


Figure 2.8: Results of the ML analysis distinguishing between precancerous and cancerous cells. From left to right: confusion matrix, ROC curves, and histogram of AUC values for (a) test data with original labels and (b) test data with randomized labels.

classifier’s performance on real and randomized data was found to be highly statistically significant, with a confidence level of $p < 0.0001$.

2.3.4 Discussion

The improved sensitivity addresses a critical clinical gap in cervical cancer screening, potentially reducing unnecessary biopsies and patient burden [40]. Although specificity is moderate, combining this AFM-ML method with existing Pap and HPV tests could yield a highly effective screening pipeline [46].

The nanoscale adhesion differences likely reflect alterations in the glycocalyx and cell surface molecular organization in cancer progression [47]. Unlike optical microscopy, AFM adhesion maps capture subtle nanoscale physical changes not visible otherwise [12, 25].

This approach is compatible with standard cytology sample preparation and may extend to other cancers, including bladder and colorectal, where similar AFM-ML methods have shown promise [35].

2.3.5 Conclusions

We demonstrate that machine learning applied to high-resolution AFM adhesion maps can robustly distinguish precancerous from cancerous cervical epithelial cells with high sensitivity and accuracy. This method holds promise for enhancing cervical cancer screening and may be adaptable to other cancer types.

Bibliography

- [1] T. Berdyeva, C. Woodworth, and I. Sokolov. “Visualization of cytoskeletal elements by the atomic force microscope”. *Ultramicroscopy* 102 (2005), pp. 189–198.
- [2] P.-H. Wu et al. “A comparison of methods to assess cell mechanical properties”. *Nature methods* 15 (2018), pp. 491–498.
- [3] I. Sokolov et al. “Noninvasive diagnostic imaging using machine-learning analysis of nanoresolution images of cell surfaces: Detection of bladder cancer”. *Proceedings of the National Academy of Sciences* 115 (2018), pp. 12920–12925.
- [4] M. Plodinec et al. “The nanomechanical signature of breast cancer”. *Nature nanotechnology* 7 (2012), pp. 757–765.
- [5] S. Iyer et al. “Atomic force microscopy detects differences in the surface brush of normal and cancerous cells”. *Nature nanotechnology* 4 (2009), pp. 389–393.
- [6] M. E. Dokukin and I. Sokolov. “Nanoscale compositional mapping of cells, tissues, and polymers with ringing mode of atomic force microscopy”. *Scientific reports* 7 (2017), p. 11828.
- [7] J. Dumont et al. “Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives”. *Critical reviews in biotechnology* 36 (2016), pp. 1110–1122.
- [8] P. Echols-Jones, W. Messner, and I. Sokolov. “Acceleration of imaging in atomic force microscopy working in sub-resonance tapping mode”. *Review of Scientific Instruments* 93 (2022).
- [9] L. Zhou et al. “Progress in the correlative atomic force microscopy and optical microscopy”. *Sensors* 17 (2017), p. 938.
- [10] Bruker Corporation. *HarmoniX Nanoscale Material Property Mapping*. <https://www.nanophys.kth.se/nanolab/afm/icon/bruker-help/Content/HarmoniX.htm>. Accessed: 2025-04-29. 2011.
- [11] N. V. Guz et al. “Towards early detection of cervical cancer: Fractal dimension of AFM images of human cervical epithelial cells at different stages of progression to cancer”. *Nanomedicine-Nanotechnology Biology and Medicine* 11 (2015), pp. 1667–1675.
- [12] M. E. Dokukin et al. “Cell Surface as a Fractal: Normal and Cancerous Cervical Cells Demonstrate Different Fractal Behavior of Surface Adhesion Maps at the Nanoscale”. *Physical review letters* 107 (2011), p. 028101.
- [13] Bruker Corporation. *PeakForce QNM*. <https://www.bruker.com/en/products-and-solutions/microscopes/materials-afm/afm-modes/peakforce-qnm.html>. Accessed: 2025-04-29. 2022.
- [14] Wikipedia contributors. *Curse of Dimensionality*. Accessed: 2025-04-29. 2025. URL: https://en.wikipedia.org/wiki/Curse_of_dimensionality.
- [15] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

-
- [16] J. Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
 - [17] Image Metrology A/S. *SPIP User’s Guide: Roughness Parameters*. Accessed: 2025-04-29. 2025. URL: https://www.imagemet.com/WebHelp6/Default.htm#RoughnessParameters/Roughness_Parameters.htm#Roughness_Average.
 - [18] R. Leach, ed. *Characterisation of Areal Surface Texture*. 2nd ed. Springer Cham, 2024.
 - [19] Digital Surf. *MountainsSPIP®*. Surface imaging and analysis software for scanning probe microscopy. Available at: <https://www.digitalsurf.com/software-solutions/scanning-probe-microscopy/> (accessed May 2025). 2025.
 - [20] J. Shlens. “A Tutorial on Principal Component Analysis”. *arXiv preprint arXiv:1404.1100* (2014).
 - [21] I. T. Jolliffe. *Principal Component Analysis*. 2nd. Springer, 2002.
 - [22] Y.-C. Tai. “A Survey Of Regression Algorithms And Connections With Deep Learning”. *arXiv preprint arXiv:2104.12647* (2021).
 - [23] L. Breiman. “Random forests”. *Machine Learning* 45 (2001), pp. 5–32.
 - [24] C. Cortes and V. Vapnik. “Support-vector networks”. *Machine Learning* 20 (1995), pp. 273–297.
 - [25] M. Dokukin et al. “Emergence of fractal geometry on the surface of human cervical epithelial cells during progression towards cancer”. *New journal of physics* 17 (2015), p. 033019.
 - [26] J.-Y. Sul et al. “Transcriptome transfer produces a predictable cellular phenotype”. *Proceedings of the National Academy of Sciences* 106 (2009), pp. 7624–7629.
 - [27] J. E. Chandler et al. “Colocalization of cellular nanostructure using confocal fluorescence and partial wave spectroscopy”. *Journal of biophotonics* 10 (2017), pp. 377–384.
 - [28] Y. Yang et al. “Transurethral partial cystectomy with 2 μ m thulium continuous wave laser in the treatment of bladder pheochromocytoma”. *Journal of endourology* 26 (2012), pp. 686–690.
 - [29] G. Mandair et al. “Raman microscopy of bladder cancer cells expressing green fluorescent protein”. *Journal of Biomedical Optics* 21 (2016), p. 115001.
 - [30] M. Dokukin et al. “Pericellular brush and mechanics of guinea pig fibroblast cells studied with AFM”. *Biophysical Journal* 111 (2016), pp. 236–246.
 - [31] M. Lekka et al. “Elasticity of normal and cancerous human bladder cells studied by scanning force microscopy”. *European Biophysics Journal* 28 (1999), pp. 312–316.
 - [32] D. B. Peckys, U. Korf, and N. de Jonge. “Local variations of HER2 dimerization in breast cancer cells discovered by correlative fluorescence and liquid electron microscopy”. *Science Advances* 1 (2015), e1500165.

- [33] M. Kijanka et al. “A novel immuno-gold labeling protocol for nanobody-based detection of HER2 in breast cancer cells using immuno-electron microscopy”. *Journal of Structural Biology* 199 (2017), pp. 1–11.
- [34] M. Petrov et al. “Detection of Human Bladder Epithelial Cancerous Cells with Atomic Force Microscopy and Machine Learning”. *Cells* 14 (2024), p. 14.
- [35] S. Prasad et al. “Atomic force microscopy detects the difference in cancer cells of different neoplastic aggressiveness via machine learning”. *Advanced NanoBiomed Research* 1 (2021), p. 2000116.
- [36] M. Petrov and I. Sokolov. “Identification of geometrical features of cell surface responsible for cancer aggressiveness: Machine learning analysis of atomic force microscopy images of human colorectal epithelial cells”. *Biomedicines* 11 (2023), p. 191.
- [37] B. Zhou et al. *Learning Deep Features for Discriminative Localization*. 2015.
- [38] E. L. Franco, N. F. Schlecht, and D. Saslow. “The epidemiology of cervical cancer”. *Cancer Journal* 9 (2003), pp. 348–359.
- [39] D. Saslow, C. D. Runowicz, D. Solomon, et al. “American Cancer Society guideline for the early detection of cervical neoplasia and cancer”. *CA: A Cancer Journal for Clinicians* 52 (2002), pp. 342–362.
- [40] M. Origoni, F. Cantatore, F. Sopracordevole, et al. “Colposcopy accuracy and diagnostic performance in Italian tertiary institutions”. *Diagnostics* 13 (2023), p. 1906.
- [41] M. Petrov and I. Sokolov. “Machine learning allows for distinguishing precancerous and cancerous human epithelial cervical cells using high-resolution AFM imaging of adhesion maps”. *Cells* 12 (2023), p. 2536.
- [42] C. Woodworth et al. “Characterization of normal human exocervical epithelial cells immortalized in vitro by papillomavirus types 16 and 18 DNA”. *Cancer research* 48 (1988), pp. 4620–4628.
- [43] I. Sokolov. *Data for the surface parameters for paper M E Dokukin et al 2015 New J. Phys.* 17 033019 DOI 10.1088/1367-2630/17/3/033019. Figshare. Dataset, posted 2023-07-31. 2023.
- [44] O. Sahin, S. Magonov, C. Su, et al. “An atomic force microscope tip designed to measure time-varying nanomechanical forces”. *Nature Nanotechnology* 2 (2007), pp. 507–514.
- [45] A. Rutherford. *ANOVA and ANCOVA: A GLM approach*. 2nd. Hoboken, N.J.: Wiley, 2011.
- [46] H. A. Katki and N. Wentzensen. “How might HPV testing be integrated into cervical screening?” *The Lancet Oncology* 13 (2012), pp. 8–10.
- [47] N. Makarova et al. “Mechanical characterization of pericellular molecular structure”. *ACS Applied Materials & Interfaces* 15 (2023), pp. 35962–35972.

Chapter 3

Mechanical spectroscopy of materials using atomic force microscopy (AFM-MS)

3.1 Introduction

Identifying material composition at the nanoscale remains a critical challenge in materials science, with major implications for the development of advanced nanocomposites and the study of microplastic contamination [1–5]. Understanding the structure-property relationships at this scale is essential for innovation in composite materials [6–10]. The proliferation of polymeric materials has also made nanoscale plastic identification important for environmental and health studies [11–14].

Traditional spectroscopic techniques, such as infrared (IR) spectroscopy [15, 16], Raman spectroscopy [17], X-ray photoelectron spectroscopy (XPS) [18], and micro-NMR [19], have been the cornerstone for material identification, but are fundamentally limited in spatial resolution. Optical spectroscopies are diffraction-limited, while XPS, despite excellent depth resolution, lacks lateral resolution. Recent advances, such as Atomic Force Microscopy-Infrared (AFM-IR) [20–23], have pushed IR spectroscopy to the nanoscale, but are restricted to thin films and ambient conditions, and are not suitable for aqueous

environments. Tip-enhanced Raman spectroscopy (TERS) [24–28] achieves high spatial resolution but is mainly limited to isolated structures and complex setups, and both AFM-IR and TERS risk sample damage from laser radiation.

Recently, there has been a lot of work done on enhancing AFM with ML for materials research [29]. In this work, we introduce a novel AFM mechano-spectroscopy (AFM-MS) technique that overcomes these limitations by combining AFM in sub-resonance tapping mode [30–34] with machine learning (ML) classification. This approach enables the collection of multiple physical and mechanical property maps with sub-nanometer lateral resolution in a highly repeatable manner. By comparing these properties to a database of known materials, the technique identifies the location of constituent materials at each image pixel with the assistance of ML algorithms. We demonstrate AFM-MS on various material mixtures, achieving an unprecedented lateral spectroscopic resolution of 1.6 nm [35]. This opens new avenues for nanoscale material study, including material identification and correlation of nanostructure with macroscopic properties.

3.2 Methods

AFM-MS utilizes Bruker Icon and Bioscope Catalyst atomic force microscopes operating in *PeakForceQNM*TM [36] mode with a Ringing mode (RM) extension [31]. ScanAsyst-Air probes (spring constant 0.3–0.4 N/m, radius 3–4 nm) were used. Imaging was performed in air at 30–40% relative humidity, with a PeakForce of 1–2 nN and a tapping rate of 1 kHz. Twelve physical and mechanical property channels were recorded simultaneously, including adhesion, deformation, dissipation, restored adhesion, viscoelastic adhesion, neck height, disconnection length, and energy loss during probe detachment.

Two distinct approaches are introduced for training ML algorithms on AFM data, differing primarily in the source of their training databases. The methods are schematically presented in Fig. 3.1, and their practical requirements are compared in Table 3.1. Method I utilizes AFM images of pure polymers, enabling the construction of a comprehensive database of nanomechanical properties that can be applied broadly, including

Requirement	Method I	Method II
Knowledge of individual (polymeric) components of the sample	Not required (optional)	Required
The external database of components (polymers)	Required	Not required
AFM imaging under a controlled environment	Frequently required	Not required
AFM imaging is restricted to the same imaging conditions as used in the external material (polymer) database	Required	Not required
Sample preparation is limited to the one used for the external material database	Required	Not required

Table 3.1: Comparison of two AFM-MS approaches for identification of components in a sample. Method I uses an existent database of nanomechanical properties of pure components (polymers). Method II uses the data collected on the part of known locations of particular pure components (polymers) within the AFM image of the study.

to samples containing unknown polymers. In contrast, Method II relies on information from known regions within the composite samples themselves, making it less generic but offering other advantages. Once training is complete, the subsequent workflow is identical: the classifier is evaluated independently for *every single pixel* in the image, thereby yielding a high-resolution spectroscopic map of the sample surface. Formally, an $N \times N$ scan contains K simultaneously acquired channels so that each pixel (i, j) is represented by a feature vector

$$\mathbf{z}_{i,j} = (z_{i,j}^{(1)}, z_{i,j}^{(2)}, \dots, z_{i,j}^{(K)}),$$

with $K = 12$ in the present work (four PeakForce-QNM and eight Ringing-Mode channels). The trained mapping CL returns a probability vector

$$\mathbf{P}_{i,j} = \text{CL}(\mathbf{z}_{i,j}) = (P_{\text{PS}}, P_{\text{PVP}}, P_{\text{PEO}})_{i,j},$$

so that polymer identity (PS, PVP and PVO in the case of a mixture of three polymers considered in this study) or class-membership likelihood is available for every coordi-

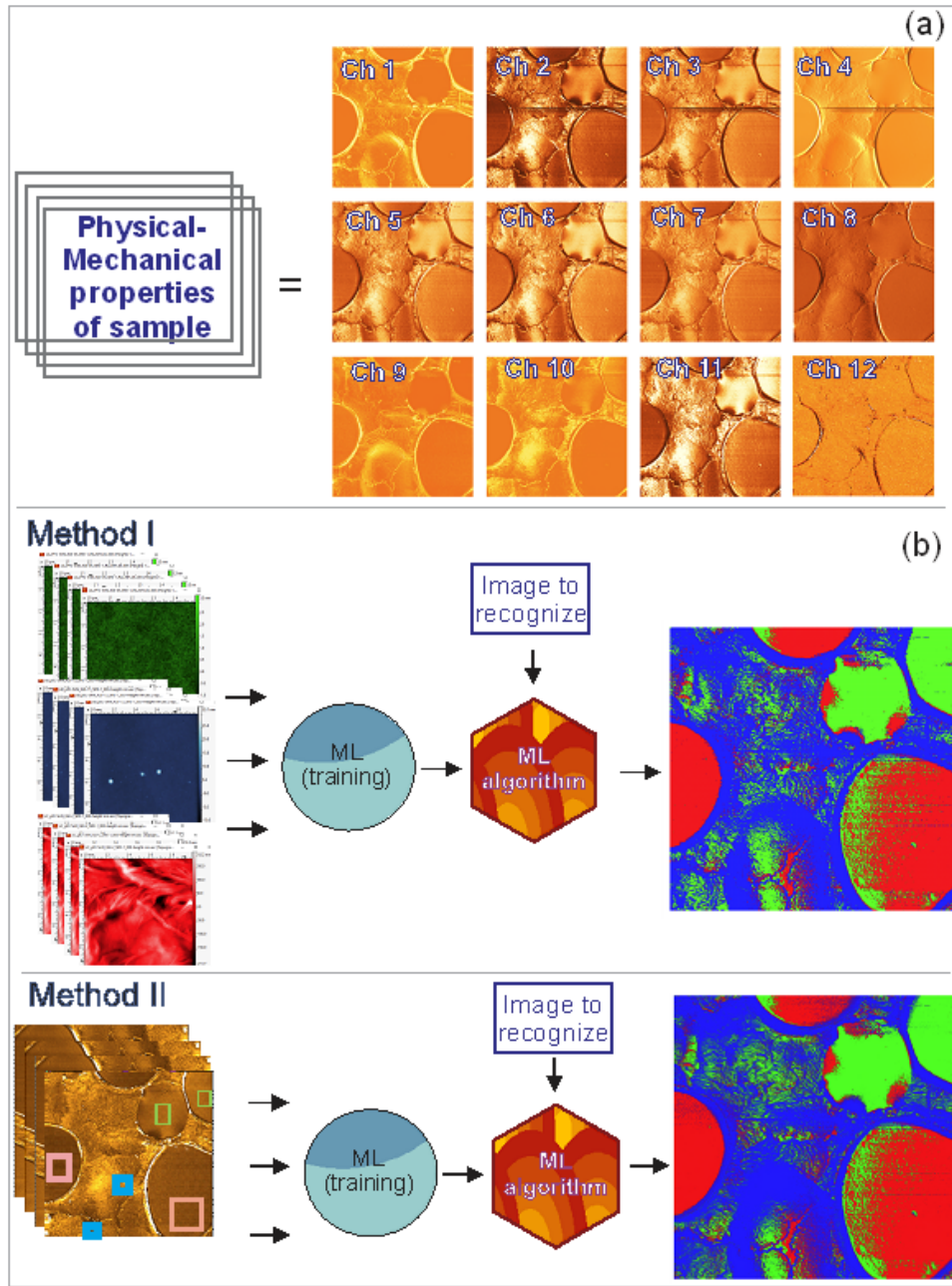


Figure 3.1: A schematic of the nano mechano-spectroscopy. (a) Physical and mechanical properties of the samples are imaged simultaneously using 12 different imaging channels. An example of a $5 \times 5 \text{ mm}^2$ blend of three different polymers, polystyrene (PS), polyethylene oxide (PEO), and polyvinyl pyrrolidone (PVP) is presented. (b) Two methods of material recognition are shown. Method I: The training of the machine learning algorithm is done using the AFM images of the corresponding pure polymers prepared in the same conditions. Method II: The training of the machine learning algorithm is done using the areas of known polymers within the sample of study (shown on the image with rectangles of different colors). Note that the AFM data can be used directly for the training of ML algorithms without any preprocessing. An example of the application of both methods to the blend of three different polymers is shown. Three different colors were used to identify particular polymers: PS is red, PVP is green, and PEO is blue. Two machine learning algorithms, (shallow) NN and RF, were used here and showed identical results.

nate in the scan. A colour-coded rendering of $\mathbf{P}_{i,j}$ produces an RGB classification map representation.

The underlying classifiers were either a Random-Forest ensemble (RF) or a shallow feed-forward neural network (NN). After empirical hyper-parameter search, the final network employed two fully connected hidden layers with seven and five neurons, respectively, followed by a soft-max output layer whose dimension matches the number of polymer classes. Pixel-wise inference on a 512×512 image therefore involves 2.6×10^5 independent forward passes, yet remains computationally tractable because the network is deliberately compact. For the RF implementation the same input vectors are classified by an ensemble of decision trees; in practice both models yielded nearly identical probability fields, confirming that the mechanically rich AFM feature space is separable without deep architectures. In both training paradigms the data set was partitioned at the scan or sub-scan level into 80% training and 20% validation subsets before the final model was frozen and applied to unseen images or image areas. Fig. 3.1 shows how the two methods applied to the same three polymer mixture sample produce identical results.

3.3 Results and Discussion

The performance of the trained ML algorithms for material recognition was evaluated using two samples composed of the same three polymers: polyethylene oxide (PEO), polyvinylpyrrolidone (PVP), and polystyrene (PS). One sample was freshly prepared, while the other was aged for two weeks. For each pixel in the AFM image, the algorithm calculated the probability that it belonged to each polymer type. The resulting probability maps for each polymer are shown individually, and an RGB composite image was generated where red, green, and blue channels correspond to PS, PVP, and PEO, respectively.

The freshly prepared sample exhibited clear spatial separation between the three polymers (Fig. 3.2 (a)), while the aged sample demonstrated significant inter-diffusion, particularly between the hydrophilic PVP and PEO polymers, likely due to environmental

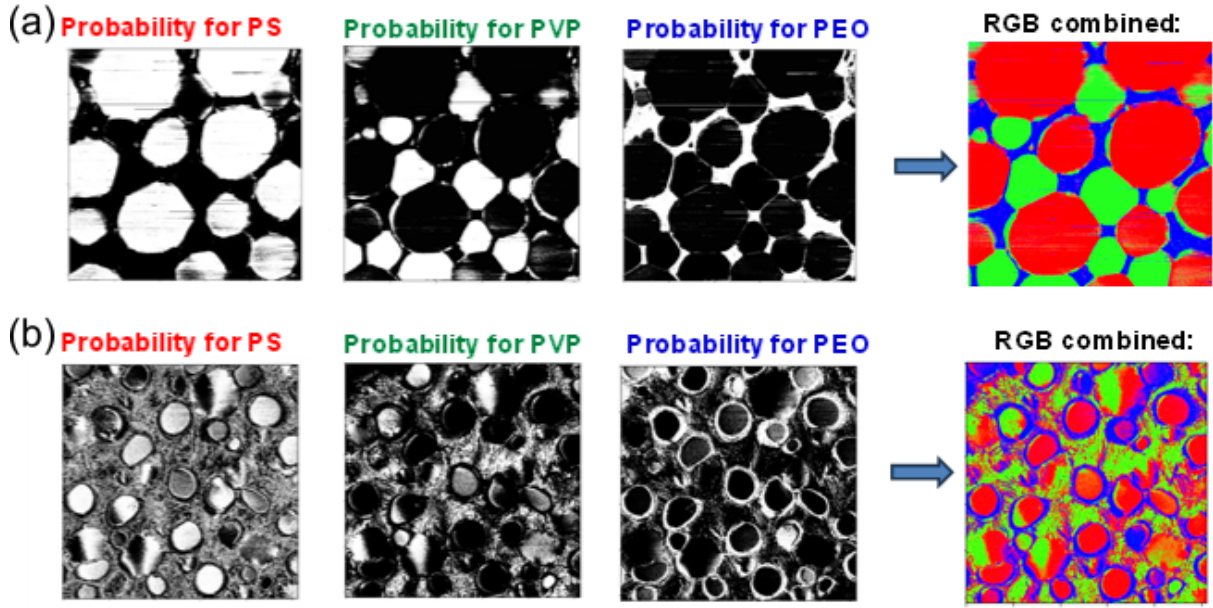


Figure 3.2: Details of the work of machine learning classification algorithms. The trained ML algorithm calculates the probability of each pixel representing a particular material. RGB combined images: The probabilities are encoded by RGB colors and combined to present the distribution of particular polymers over the sample. Two samples of $20 \times 20 \mu\text{m}^2$ are shown. (a) A freshly prepared mix of three polymers (PS, PVP, PEO). (b) The same sample two weeks after the preparation; polymer diffusion makes the sample more challenging for recognition.

humidity during storage (Fig. 3.2 (b)). This observation highlights the capability of the presented technique to study surface diffusion and mixing phenomena at the nanoscale.

The lateral spectroscopic resolution of the technique is fundamentally determined by the contact area of a single AFM indentation, which depends on the probe radius, load force, adhesion energy, and indentation speed. In this work, the AFM probe had a radius of curvature of approximately 2–3 nm, and the indentation depth did not exceed 2–4 nm, suggesting that the achievable lateral resolution is on the order of a few nanometers. However, while the theoretical spectroscopic resolution is defined by the pixel size, practical resolution may be limited by variations in measured properties within each material. Although visual inspection of single-channel images allows some discrimination between the polymers, histograms of physical property distributions reveal substantial overlap, indicating that multi-channel analysis is essential for reliable identification.

The technique’s spectroscopic resolution was further demonstrated using mechanically distinct materials. For example, Fig. 3.3 shows that the method can resolve a silica nanoparticle embedded in epoxy resin with a lateral resolution of approximately 1.6 nm,

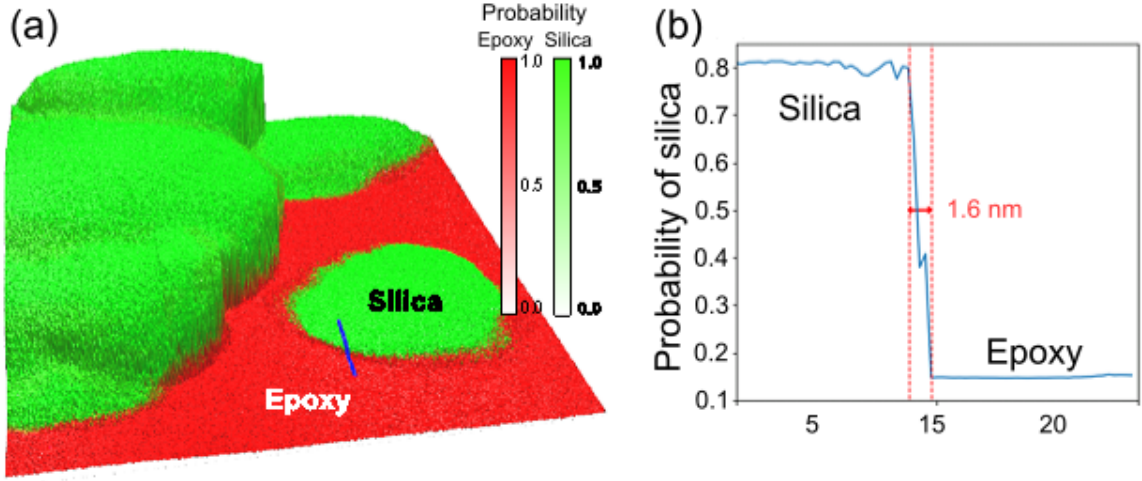


Figure 3.3: Experimental demonstration of the lateral spectral resolution of the described AFM-MS. (a) 3D height AFM image of $200 \times 200 \text{ nm}^2$ is shown. The probabilities of identification of material as epoxy (red) and silica (green) are superimposed on the height image using different colors. (b) A cross-section of the probability of identifying material as silica is shown. The specific location of this cross-section is shown in panel (a) by a solid blue line. The width of the probability change of 1.6 nm, which can be called the lateral spectral resolution [21], is shown.

consistent with theoretical expectations and limited primarily by the AFM probe’s radius.

A comparison with confocal Raman spectroscopy (Fig. 3.4), a widely used material identification technique, underscores the advantages of AFM-MS. Raman microscopy is limited by the optical diffraction limit, resulting in a lateral resolution of about 200 nm, and its vertical resolution is further constrained by the sample’s thickness and optical parameters, typically around $0.5 \mu\text{m}$. In contrast, the nano-mechano-spectroscopy technique achieves both lateral and vertical resolutions on the nanometer scale, determined by the AFM probe’s interaction with the sample surface. However, it is limited to surface analysis and cannot detect materials buried beneath the surface, unlike Raman spectroscopy.

The complementary nature of these two techniques is noteworthy. Raman spectroscopy can be used to identify large regions of a sample, providing valuable training data for the machine learning algorithms employed in Method II. Subsequently, nano-mechano-spectroscopy can pinpoint the precise surface locations of different materials with nanoscale accuracy, offering a powerful combination for comprehensive material characterization.

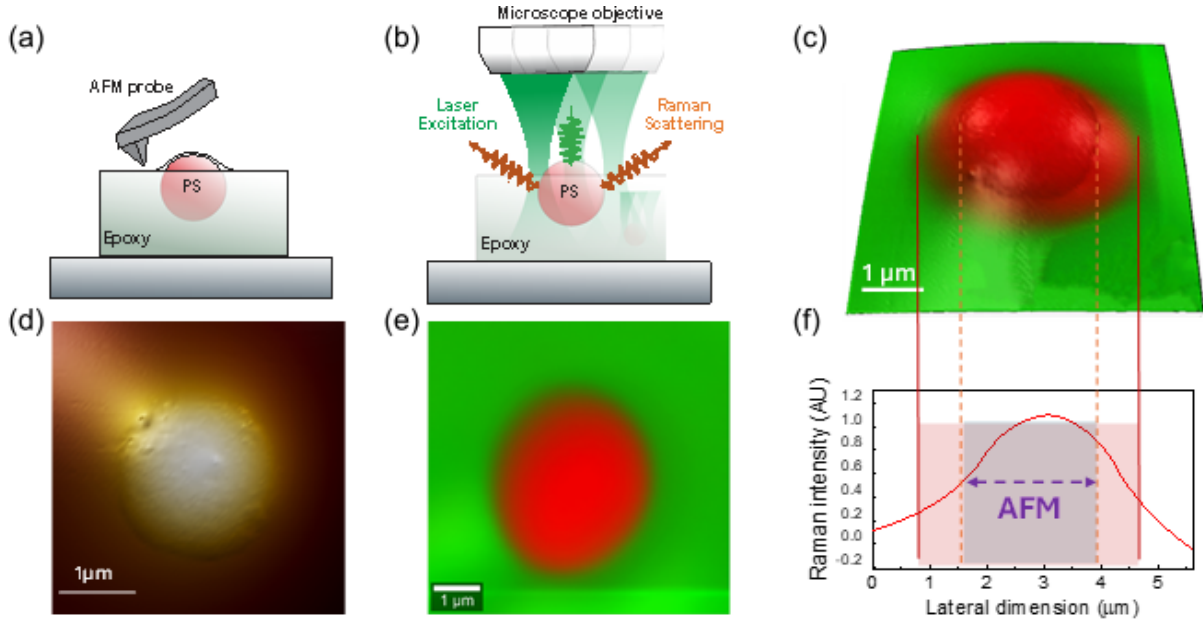


Figure 3.4: Comparison with confocal Raman microscopy/spectroscopy and mechano-spectroscopy: an example of a polystyrene sphere partially embedded in epoxy. (a) A schematic of the sample. The AFM probe can easily image the outstanding part of the polystyrene bead. Substantial differences in roughness and adhesion (not shown) allow easy identification of the bead above the epoxy. (b) A diagram showing the work of Raman microscopy. The finite depth of focus of the laser excitation light may result in a Raman signal coming from both polystyrene and epoxy. (c) A 3D AFM image of the polystyrene sphere embedded in epoxy (shown in panel d) superimposed with the map of the Raman signal (shown in panel e). (f) Shows the size of the outstanding part of the polystyrene bead as well as the location of the detectable Raman signal associated with polystyrene.

3.4 Conclusion

This study presents a novel AFM-MS technique that leverages the physical and mechanical properties of materials to identify and spatially resolve specific components within composite samples. Achieving single-nanometer spatial resolution, this approach surpasses the limitations of conventional material identification methods and offers significant advantages for the study and design of complex, nanostructured materials. Its applicability spans polymers, biomaterials, nanocomposites, and other advanced materials, providing unprecedented detail for both fundamental research and engineering applications. Furthermore, the unique information obtained through this technique can be synergistically integrated with complementary methods, such as confocal Raman microscopy, to enable comprehensive, multi-modal material characterization. The ability to

map material composition and properties with such high precision is poised to advance our understanding and optimization of next-generation materials.

Bibliography

- [1] U. Scherf and E. J. List. “Semiconducting polyfluorenes—towards reliable structure–property relationships”. *Advanced Materials* 14 (2002), pp. 477–487.
- [2] J. Jancar et al. “Current issues in research on structure–property relationships in polymer nanocomposites”. *Polymer* 51 (2010), pp. 3321–3343.
- [3] T. Le et al. “Quantitative structure–property relationship modeling of diverse materials properties”. *Chemical reviews* 112 (2012), pp. 2889–2919.
- [4] E. D. Cubuk et al. “Structure–property relationships from universal signatures of plasticity in disordered solids”. *Science* 358 (2017), pp. 1033–1037.
- [5] Y. Yao et al. “High-entropy nanoparticles: Synthesis–structure–property relationships and data-driven discovery”. *Science* 376 (2022), eabn3103.
- [6] Y. Zeng et al. “Irreversible synthesis of an ultrastrong two-dimensional polymeric material”. *Nature* 602 (2022), pp. 91–95.
- [7] P. A. Klonos et al. “Structure–properties relationships in new polymer nanocomposites based on the renewable poly (butylene succinate) filled with low amounts of nanoparticles of 1-3D geometries”. *Polymer* 296 (2024), p. 126841.
- [8] G. Toader et al. “State-of-the-Art Polyurea Coatings: Synthesis Aspects, Structure–Properties Relationship, and Nanocomposites for Ballistic Protection Applications”. *Polymers* 16 (2024), p. 454.
- [9] B. Wongvasana et al. “Comparative structure–property relationship between nanoclay and cellulose nanofiber reinforced natural rubber nanocomposites”. *Polymers* 14 (2022), p. 3747.
- [10] M. Zhang et al. “The Microscopic Structure–Property Relationship of Metal–Organic Polyhedron Nanocomposites”. *Angewandte Chemie* 131 (2019), pp. 17573–17578.
- [11] V. Nava et al. “Plastic debris in lakes and reservoirs”. *Nature* 619 (2023), pp. 317–322.
- [12] S. Huang et al. “Risk effects of meteorological factors on human brucellosis in Jilin province, China, 2005–2019”. *Heliyon* 10 (2024).
- [13] B. Muthusubramanian et al. “Assessing the embodied carbon and energy required for manufacturing sustainable concrete blocks using plastic pollution as a fiber”. *Environmental Science and Pollution Research* 30 (2023), pp. 107533–107548.
- [14] S. Ozden et al. “Egg protein derived ultralightweight hybrid monolithic aerogel for water purification”. *Materials Today* 59 (2022), pp. 46–55.
- [15] L. Shi et al. “Mid-infrared metabolic imaging with vibrational probes”. *Nature methods* 17 (2020), pp. 844–851.
- [16] M. J. Nasse et al. “High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams”. *Nature methods* 8 (2011), pp. 413–416.
- [17] J. Cheng and X. Xie. “Coherent Raman scattering microscopy of biological tissues”. *Science* 350 (2015), p. 6264.

-
- [18] D. N. G. Krishna and J. Philip. “Review on surface-characterization applications of X-ray photoelectron spectroscopy (XPS): Recent developments and challenges”. *Applied Surface Science Advances* 12 (2022), p. 100332.
- [19] R. D. Allert, K. D. Briegel, and D. B. Bucher. “Advances in nano-and microscale NMR spectroscopy using diamond quantum sensors”. *Chemical Communications* 58 (2022), pp. 8165–8181.
- [20] A. Dazzi and C. B. Prater. “AFM-IR: Technology and applications in nanoscale infrared spectroscopy and chemical imaging”. *Chemical reviews* 117 (2017), pp. 5146–5173.
- [21] L. Wang et al. “Nanoscale simultaneous chemical and mechanical imaging via peak force infrared microscopy”. *Science advances* 3 (2017), e1700255.
- [22] N. Piergies et al. “Nanoscale image of the drug/metal mono-layer interaction: Tapping AFM-IR investigations”. *Nano Research* 13 (2020), pp. 1020–1028.
- [23] Z. Wang et al. “Molecular orientation in individual electrospun nanofibers studied by polarized AFM-IR”. *Macromolecules* 52 (2019), pp. 9639–9645.
- [24] B.-S. Yeo et al. “Nanoscale probing of a polymer-blend thin film with tip-enhanced raman spectroscopy”. *Small* 5 (2009), pp. 952–960.
- [25] Y. Pandey et al. “Nanoscale visualization of phase separation in binary supported lipid monolayer using tip-enhanced Raman spectroscopy”. *Nanoscale* 16 (2024), pp. 10578–10583.
- [26] Z.-F. Cai et al. “Mechanistic Understanding of Oxygen Activation on Bulk Au (111) Surface Using Tip-Enhanced Raman Spectroscopy”. *Angewandte Chemie* 136 (2024), e202318682.
- [27] M. Filez et al. “Nanoscale Chemical Diversity of Coke Deposits on Nanoprinted Metal Catalysts Visualized by Tip-Enhanced Raman Spectroscopy”. *Advanced Materials* 36 (2024), p. 2305984.
- [28] E. Bailo and V. Deckert. “Tip-enhanced Raman scattering”. *Chemical Society Reviews* 37 (2008), pp. 921–930.
- [29] U. Pratiush et al. “Mic-hackathon 2024: Hackathon on Machine Learning for Electron and Scanning Probe Microscopy”. *arXiv preprint arXiv:2506.08423* (2025).
- [30] G. Smolyakov et al. “AFM PeakForce QNM mode: Evidencing nanometre-scale mechanical properties of chitin-silica hybrid nanocomposites”. *Carbohydrate polymers* 151 (2016), pp. 373–380.
- [31] M. E. Dokukin and I. Sokolov. “Nanoscale compositional mapping of cells, tissues, and polymers with ringing mode of atomic force microscopy”. *Scientific reports* 7 (2017), p. 11828.
- [32] M. E. Dokukin and I. Sokolov. “Quantitative mapping of the elastic modulus of soft materials with HarmoniX and PeakForce QNM AFM modes”. *Langmuir* 28 (2012), pp. 16060–16071.
- [33] Y. M. Efremov et al. “Viscoelastic mapping of cells based on fast force volume and PeakForce Tapping”. *Soft Matter* 15 (2019), pp. 5455–5463.

- [34] L. Li et al. “Measurement of nanomechanical properties of DNA molecules by PeakForce atomic force microscopy based on DNA origami”. *Nanoscale* 11 (2019), pp. 4707–4711.
- [35] M. Petrov et al. “Mechanical spectroscopy of materials using atomic force microscopy (AFM-MS)”. *Materials Today* 80 (2024), pp. 218–225.
- [36] Bruker Corporation. *PeakForce QNM*.
<https://www.bruker.com/en/products-and-solutions/microscopes/materials-afm/afm-modes/peakforce-qnm.html>.
Accessed: 2025-04-29. 2022.

Chapter 4

Bayesian deep transfer learning

Modern deep learning algorithms frequently contain millions of trainable parameters, performing optimally with abundant data. In practice, dataset size remains a critical constraint—particularly in medical applications where acquiring large quantities of high-quality data is both costly and logistically challenging. This problem was already discussed at length in Chapter 2 of this thesis with respect to AFM images of cells for cancer detection.

Recent advances in *transfer learning* offer promising solutions to data scarcity. This paradigm involves pre-training NNs on abundant, task-adjacent datasets before adapting them to specialized applications with limited samples (sometimes fewer than 100 observations). Such approaches effectively bypass the traditional trade-off between model complexity and dataset size.

This chapter focuses on developing such algorithms, with a future goal of benchmarking them on our AFM data.

- Section 4.1 critically analyzes a recently proposed Bayesian transfer learning framework. Unlike conventional approaches that produce a single optimized set of weights during pre-training, this method constructs a probability distribution over the weight space, enabling more robust adaptation to target tasks. A particularly relevant benchmark in this work is the application of this framework to the HAM10000 dataset, which consists of skin cancer images.

- Section 4.2 presents a hyperparameter optimization framework that serves as a general-purpose training protocol for neural networks, with demonstrated efficacy in accelerating Bayesian transfer learning pipelines.

4.1 Transfer Learning with Informative Priors

4.1.1 Introduction

Dataset size fundamentally limits the achievable performance of deep learning models, particularly in applications like medical image classification where labeled data are scarce due to the high cost and expertise required for annotation. When only a small labeled dataset is available—often just a few dozen or hundred examples—models are prone to overfitting and poor generalization, making it critical to seek alternative sources of information to enhance learning [1]. Transfer learning addresses this challenge by leveraging knowledge from a large labeled “source” dataset, such as ImageNet [2], or a pre-trained model derived from such data [3]. The standard transfer learning workflow involves two steps: first, pre-training a neural network on the source data to obtain a point estimate of the network weights; second, using these weights to initialize the model for supervised training on the smaller “target” dataset, a process known as fine-tuning [4, Sec. 19.2]. This approach is widely adopted in medical imaging, where it enables effective model training even with limited labeled data, as features learned from natural images can be adapted to medical tasks. Recent advances, such as Bayesian transfer learning BTL, propose further improvements by learning a distribution over network weights from the source task rather than a single point estimate [5]. This distribution serves as an informative prior for the target task. For instance, in classification tasks with limited data (e.g., 1000 samples on CIFAR-10), Bayesian methods using maximum a-posteriori (MAP) estimation with informative priors have demonstrated significant reductions in error rates compared to standard transfer learning. Additional gains have been reported when estimating the full posterior distribution over weights via techniques like stochastic gradient Hamiltonian Monte Carlo (SGHMC) [6]. However, our empirical investigations

reveal that the relative advantage of MAP-based transfer learning with informative priors over standard transfer learning can vary substantially across datasets. In systematic experiments spanning natural and medical image datasets, standard transfer learning often performs better than previously reported, and the magnitude of gains from Bayesian approaches depends on the specific dataset and task. Thus, while advanced transfer learning methods hold promise, their benefits are context-dependent, and standard transfer learning remains a robust baseline, especially in medical imaging where labeled data are limited. Our inability to reproduce previously reported results is part of a larger trend across machine learning research [7].

4.1.2 Methodology

Bayesian transfer learning.

A recent work proposed BTL method, where a re-scaled posterior from the source task is used as the prior for the target task [5]. This work is motivated by sequential Bayesian updating, where some source data $\mathcal{D}_S$ is acquired, a posterior $p(w|\mathcal{D}_S)$ over weights w is formed, and this posterior is used as an *informed prior* for a target dataset $\mathcal{D}_T$ [8, Sec. 3.3.3]

$$p(w|\mathcal{D}_S) \propto p(\mathcal{D}_S|w)p(w) \tag{4.1}$$

$$p(w|\mathcal{D}_T, \mathcal{D}_S) \propto p(\mathcal{D}_T|w)p(w|\mathcal{D}_S).$$

In sequential Bayesian updating, it is typically assumed that all data—once conditioned on the model parameters—are independent and identically distributed. However, in transfer learning, this assumption does not strictly hold because the target data are drawn from a distribution that may differ from the source data. As a result, the prior beliefs about which parameter values are plausible for the target task may require adjustment. To address this, a scaling factor λ is introduced to control the influence of the source-informed prior over the weights w . Specifically, λ rescales the covariance matrix of

the posterior $p(w|\mathcal{D}_S)$ (which is assumed to be a Gaussian) to form the source-informed prior. When $\lambda = 1$, the source posterior is used directly as the prior for the target, treating the source and target as if they were identically distributed. As λ increases, the covariance becomes more diffuse, weakening the influence of the source data. In the limit as $\lambda \rightarrow \infty$, the prior becomes effectively uniform, and the only remaining benefit from the source task is the initialization of the model parameters—thereby recovering standard transfer learning.

Common framework for MAP transfer learning

Probabilistic model for target task. Building on the previously introduced framework [5], we construct a model for image classification with C possible classes. The architecture consists of two main components: a feature extraction backbone and a classification head. Let f denote the backbone neural network, parameterized by weights $w \in \mathbb{R}^d$, where d may range from millions to billions in modern computer vision models. For any input image x_i of fixed size, the backbone produces a hidden representation $f_w(x_i) \in \mathbb{R}^H$. We further assume that the first entry of this representation is fixed to 1, allowing the model to learn class-specific intercepts. The classifier head is parameterized by a weight matrix $V \in \mathbb{R}^{C \times H}$, which maps the hidden representation to a vector of logits, one for each class. Given a target training set of labeled examples, $\{x_i, y_i\}_{i=1}^n$, where each x_i is an image and $y_i \in \{1, 2, \dots, C\}$ is its class label, the overall probabilistic model can be summarized as follows:

- The backbone transforms each input image into a feature vector.
- The classifier head produces logits from these features, which are converted into class probabilities via the softmax function.
- The model is trained to maximize the likelihood of the observed labels under this probabilistic mapping.

Formally, the model can be described by the following components:

$$p(w) = \mathcal{N}(w \mid \mathbf{m}, \lambda \mathbf{S}) \quad \text{source-informed prior on backbone} \quad (4.2)$$

$$p(V) = \mathcal{N}(\text{vec}(V) \mid 0, \tau I), \quad \text{prior on clf. head} \quad (4.3)$$

$$p(y_{1:n} \mid x_{1:n}, w, V) = \prod_{i=1}^n \text{Cat}(y_i \mid \mathcal{S}(V f_w(x_i))) \quad \text{likelihood} \quad (4.4)$$

Here, $\mathcal{S}$ denotes the softmax operation, which converts logits into probabilities over the classes y_i . $\tau > 0$ and $\lambda > 0$ are key hyperparameters that require tuning, while the choice of $\mathbf{m}, \mathbf{S}$ is method specific; these values may be simple (e.g. an all-zero mean) or informed by a source-task.

Target task MAP estimation. To fit the probabilistic model described above to a target dataset, we employ a MAP estimation strategy. This involves finding the values of the backbone and classifier head weights, w and V , that minimize the following objective function:

$$L(w, V) := -\frac{1}{n} \left[\sum_{i=1}^n \log p(y_i \mid x_i, w, V) + \log p(w) + \log p(V) \right] \quad (4.5)$$

Here, the first term encourages the model to assign high likelihood to the observed labels, while the second and third terms incorporate prior knowledge about the weights, acting as regularizers. For optimization, we use stochastic first-order gradient descent (SGD), a standard approach for large-scale neural network training. In the context of transfer learning, it is customary to initialize the backbone weights w with values μ obtained from pre-training on the source task, leveraging previously acquired knowledge. The classifier head weights V are typically initialized randomly, as they are specific to the target task.

Various transfer learning approaches in the literature adopt this common modeling and MAP estimation framework. The primary differences among them lie in the choice of the prior mean vector $\mathbf{m}$ and covariance matrix $\mathbf{S}$ for the backbone weights. Table 4.1 provides an overview of specific methods and their corresponding prior configurations.

Method	also known as	Prior	Init.
StdPrior fromScratch	SGD Non-Learned Prior [5]	$\mathcal{N}(0, \lambda I)$	random
StdPrior fromImgNet	SGD Transfer Init [5]	$\mathcal{N}(0, \lambda I)$	μ
LearnedPriorIso fromImgNet	MAP adaptation [9]	$\mathcal{N}(\mu, \lambda I)$	μ
LearnedPriorLR fromImgNet	SGD Learned Prior [5]	$\mathcal{N}(\mu, \lambda \Sigma)$	μ

Table 4.1: Possible methods for point estimation of neural network weights for a target task.

Experimental setup

In this work, we conducted a rigorous and reproducible comparison of three transfer learning methods for point estimation of neural network weights on a target image classification task. Each method leverages pre-training on a source task, either through initialization alone or by incorporating a learned prior. These approaches are

- *StdPrior*, which uses a standard Gaussian prior with zero mean and isotropic covariance
- *LearnedPriorIso*, which employs a source-informed prior with a learned mean but retains isotropic covariance
- *LearnedPriorLR*, which utilizes both a learned mean and a LR covariance structure derived from the source task

We evaluate these methods on five open-access datasets spanning a range of classification challenges: CIFAR-10 [10], Oxford Flowers [11], Oxford-IIIT Pets [12], FGVC-Aircraft [13], and HAM10000 [14, 15]. Each dataset presents unique characteristics, from general object recognition to fine-grained and medical image classification. For all datasets, we use a ResNet-50 [16] backbone and adhere to a consistent optimization protocol, employing SGD with Nesterov momentum and a cosine-annealing learning rate schedule [17]. For all methods, we took the learned weights μ that form the initialization (and if needed, the prior mean), from publicly released snapshots ¹ [5]. Transfer learning methods are particularly valuable in low-data regimes, so for each dataset, we evaluated

¹ SimCLR snapshots [5] can be found at <https://github.com/hsouri/BayesianTransferLearning>

multiple training set sizes, carefully selected from the original data in a controlled and reproducible manner (see details Appendix B.2).

4.1.3 Results

The results of the conducted experiments can be summarized in three main findings.

First, standard transfer learning better than reported. The comparison between current experiments and prior work [5] reveals significant improvements in standard transfer learning performance when using initialization-only approaches. Fig. 4.1 demonstrates that the *StdPrior* method achieves error rates 10-15 percentage points lower than previously reported at dataset sizes of $n = 100$ and $n = 1000$ examples, effectively eliminating the performance gap that had been attributed to more complex informative prior methods. This finding is corroborated by independent experiments [18], whose ResNet-50/ImageNet implementation shows test errors of 17-22% at $n=1000$ training examples, closely matching the current results. The consistency across studies suggests standard transfer learning can achieve substantially better results than earlier benchmarks suggested.

Second, relative gains of informed priors over standard transfer learning vary across datasets. The ultimate test set performance for all evaluated methods across varying train set sizes is detailed for five datasets, with each table entry showing the mean test accuracy and the range across three replicate samples (Tables 4.2, 4.3, 4.4, 4.5 and 4.6). The benefit of using informative priors over standard transfer learning is inconsistent and depends on the dataset: for Oxford Flowers and HAM10000, informative priors offer little to no improvement; on CIFAR-10 and Oxford-IIIT Pets, the gains are modest (1.5-3 percentage points); and on FGVC-Aircraft, the gains are substantial, exceeding eight points. Among the informative prior methods, isotropic covariance [9] performs comparably to more complex LR covariance [5], despite being simpler to implement and tune. While there are a few instances where the LR approach outperforms the isotropic one—such as on CIFAR-10 and Oxford-IIIT Pets at certain training set sizes—these advantages are not consistent across all datasets, and in cases of significant improvement,

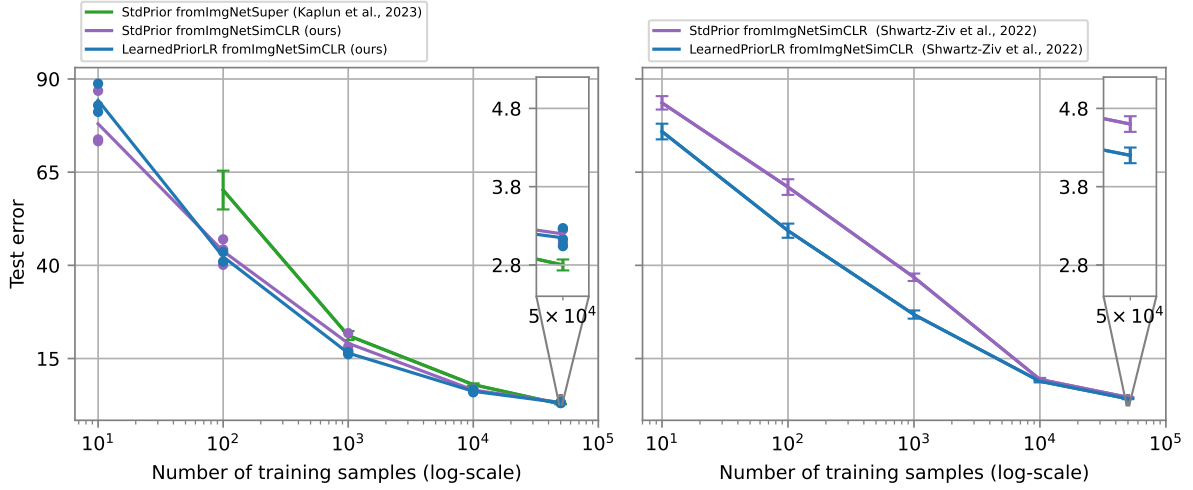


Figure 4.1: Error rate (lower is better) vs. target train set size on CIFAR-10, for various MAP estimation methods for transfer learning from ImageNet. *Left:* Our results. *Right:* Results copied from [5] (their Tab. 10). *Setting details:* The blue and purple lines across both panels come from comparable settings: a common ResNet-50 architecture and common learned values for mean and low-rank (LR) covariance taken directly from the SimCLR pre-trained snapshots in [5]’s repository. *Green line:* The left panel’s green line is a third-party experiment copied from [18].

the simpler isotropic method remains competitive.

Third, large variability in quality of alignment between training and test loss landscapes. Previous work propose that informed priors enhance transfer learning by improving the alignment between the loss landscapes of the target task’s small training set and its large test set, ideally making the locations of minima more similar and thus boosting generalization [5]. Contrary to this argument, when we empirically examined—specifically by interpolating between different model optima on CIFAR-10 with $n = 1000$ training examples—the results revealed considerable variability. Our analysis for loss alignment is shown in Fig. 4.2. We assess a 1D slice of the high-dimensional landscape by linearly interpolating parameters w between the optima w_{Std}^* found via minimizing the standard MAP objective (left) or the optima w_{LR}^* found via minimizing the *LearnedPriorLR* MAP objective (right) and the optima w^* found at the largest dataset size. Red curve shows the indicated training loss (varies by column) on CIFAR-10 with $n = 1000$ samples; blue curve (varies by column) shows CIFAR-10 test set negative log-likelihood (NLL). Each row shows results from a different train set sample of size $n = 1000$. The gap between the

optima found via training and the test set’s ideal minimum is shown as a double-sided gold arrow.

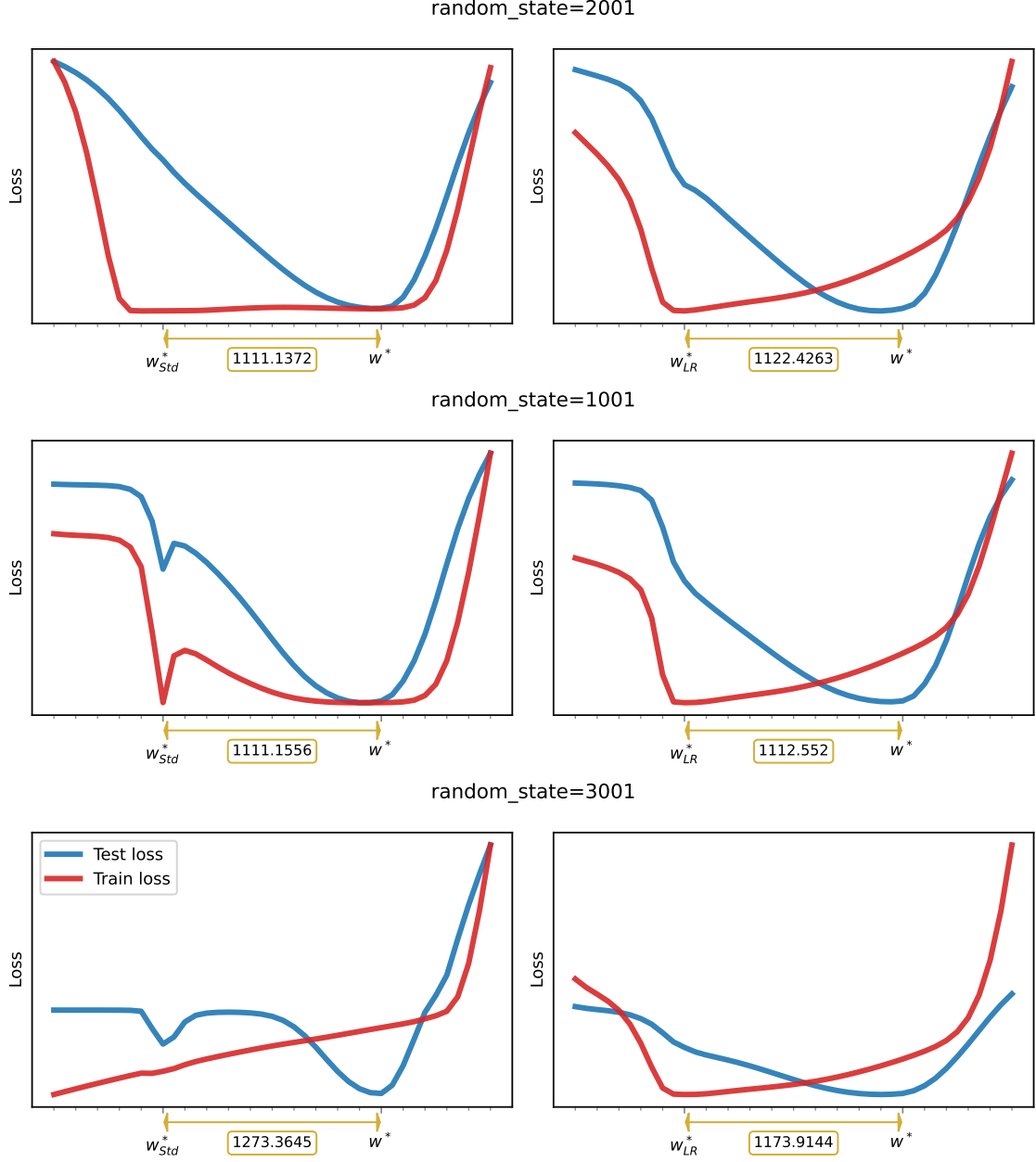


Figure 4.2: Empirical alignment of loss landscapes for target task across train and test for CIFAR-10 $n = 1000$. Compare to Shwartz-Ziv et al.’s Figure 1, which is an idealized illustration not an empirical result.

Sometimes the informed prior does bring the train and test minima closer together, but in other cases, the alignment is no better than with standard methods, or even worse.

Method	$n = 10$ (1/cl.)	100 (10/cl.)	1000 (100/cl.)	10000 (1k/cl.)	50000 (5k/cl.)
StdPrior fromImgNet	22.0 (13.2-26.7)	56.2 (53.0-59.9)	80.9 (78.2-82.8)	93.4 (93.3-93.7)	96.8 (96.7-96.9)
LearnedPriorIso fromImgNet	20.7 (17.7-23.4)	56.7 (54.8-59.8)	81.9 (81.6-82.1)	94.3 (94.1-94.5)	97.3 (97.2-97.4)
LearnedPriorLR fromImgNet	15.7 (11.3-18.8)	58.7 (56.4-60.6)	83.5 (83.3-83.9)	93.8 (93.4-94.1)	96.9 (96.7-97.0)

Table 4.2: CIFAR-10 heldout accuracy (higher is better) as target train set size n increases.

Method	$n = 102$ (1/cl.)	510 (5/cl.)	1020 (10/cl.)
StdPrior fromImgNet	31.1 (12.6-40.6)	78.9 (78.4-79.3)	87.9 (87.4-88.1)
LearnedPriorIso fromImgNet	19.2 (9.0-34.8)	79.1 (78.5-79.4)	88.6 (88.2-89.1)
LearnedPriorLR fromImgNet	28.8 (10.6-39.2)	77.9 (74.8-79.4)	88.4 (88.2-88.7)

Table 4.3: Oxford Flowers heldout accuracy (higher is better) as target train set size n increases.

Method	$n = 37$ (1/cl.)	370 (10/cl.)	3441(93/cl.)
StdPrior fromImgNet	17.3 (14.5-20.0)	54.8 (53.2-57.5)	86.4 (86.0-86.6)
LearnedPriorIso fromImgNet	7.0 (5.4- 8.5)	55.5 (53.3-57.7)	86.4 (85.4-87.0)
LearnedPriorLR fromImgNet	6.6 (6.2- 7.3)	57.4 (56.2-58.2)	86.7 (85.0-87.8)

Table 4.4: Oxford-IIIT Pets heldout accuracy (higher is better) as target train set size n increases.

Method	$n = 100$ (1/cl.)	500 (5/cl.)	1000 (10/cl.)	5000 (50/cl.)
StdPrior fromImgNet	2.6 (1.3-4.6)	22.5 (21.4-24.4)	40.6 (39.8-41.8)	85.7 (85.4-86.0)
StdPrior fromImgNet	3.2 (1.8-4.6)	22.5 (21.4-24.4)	44.1 (39.8-46.6)	85.7 (85.4-86.0)
LearnedPriorIso fromImgNet	3.5 (2.8-4.5)	25.9 (24.0-27.3)	51.3 (50.0-52.6)	85.9 (85.8-86.0)
LearnedPriorLR fromImgNet	3.8 (3.5-4.2)	24.5 (23.7-25.3)	50.9 (50.4-51.9)	84.2 (83.5-85.3)

Table 4.5: FGVC-Aircraft heldout accuracy (higher is better) as target train set size n increases.

Method	$n = 100$	1000
StdPrior fromImgNet	78.1 (75.0-82.8)	85.6 (85.0-86.3)
LearnedPriorIso fromImgNet	78.0 (75.0-82.8)	86.5 (85.5-87.4)
LearnedPriorLR fromImgNet	78.7 (74.9-82.7)	86.6 (85.0-87.5)

Table 4.6: HAM10000 AUROC on heldout data (higher is better, macro-averaged across classes) as target train set size n increases.

This inconsistency across replicate samples suggests that the empirical evidence for the loss-alignment mechanism is much weaker and more variable than reported previously [5].

4.1.4 Conclusion

The findings are limited to the simpler MAP point estimation setting and do not address the full Bayesian posterior estimation methods, which are computationally more intensive and require greater expertise. Nevertheless, standard transfer learning results on CIFAR-10 match or exceed the full Bayesian inference outcomes reported in earlier work [5], and it is anticipated that deep ensembles could further enhance performance. Future research is needed to compare full Bayesian inference using source-informed priors to well-tuned deep ensemble baselines.

For point estimation of neural network weights, both standard priors and isotropic learned priors are recommended due to their simplicity and strong performance. The LR covariance approach does not consistently provide enough improvement to justify its use as a default, though it may offer modest gains in certain small-data scenarios.

The study is limited by its focus on transfer learning from a single source task (ImageNet) and by using fixed prior definitions based on released model snapshots. Further research could explore whether informative priors become more beneficial when the source task is more closely related to the target task, or when different pre-training procedures and prior configurations are used.

The overarching aim is to advance the community’s understanding of effective transfer learning strategies. Two best practices are highlighted for future work: ensuring code is shared to enable full reproducibility of experiments, including data preparation and hyperparameter tuning, and providing equal time and computational resources for hyperparameter tuning across all evaluated methods.

4.2 Learning Hyperparameters via a Data-Emphasized Variational Objective

4.2.1 Introduction

The previous section discussed how transfer learning can benefit from using a pre-trained weight space distribution rather than relying on a single set of pre-trained weights. Although these models provide greater flexibility for adapting to downstream tasks, they introduce a significantly larger hyperparameter space. This expanded space is costly to tune, which can limit both the effectiveness and practical applicability of the approach [19]. In fact, when fine-tuning deep NN, a substantial portion of computational effort is spent tuning hyperparameters that govern model complexity, such as the scalar multiplier for L2 regularization [20] (also known as Ridge penalty [21, 22] or weight decay [23, 24]). The conventional approach for hyperparameter selection is grid search, where a dedicated validation set is used to evaluate model performance across a predefined set of hyperparameter values [25, 26]. While this method is popular and effective for mitigating overfitting [5, 27], it suffers from three major limitations: it requires training separate models for every candidate hyperparameter value, which is computationally expensive; it reduces the amount of data available for training by reserving a validation set, potentially leading to underfitting when data is scarce; and it restricts the search to a discrete set of values that must be chosen in advance, which may not include the optimal setting for a given dataset or task.

To address these drawbacks, we proposed a Bayesian approach to hyperparameter selection. In this framework, the dataset $\mathcal{D}$ is modeled with a likelihood function $p(\mathcal{D}|\theta)$ parameterized by model parameters θ , which themselves have a prior distribution $p(\theta|\eta)$ controlled by a hyperparameter η . Instead of selecting a single best θ, η pair of model parameters and hyperparameters, the Bayesian method estimates the posterior distribution $p(\theta|\mathcal{D}, \eta)$ over model parameters for a given hyperparameter setting, and then seeks to optimize the marginal likelihood (or evidence) $p(\mathcal{D}|\eta) = \int_{\theta} p(\mathcal{D}, \theta|\eta) d\theta$ with respect to the

η [28]. Maximizing the evidence naturally incorporates Occam’s razor, favoring simpler models that fit the data well and penalizing overly complex models that risk overfitting.

Optimizing η by maximizing the evidence, typically via gradient descent, offers several advantages over grid search: it requires only a single training run rather than multiple runs for each hyperparameter candidate, it allows all available labeled data to be used for training without the need for a separate validation set, and it enables exploration of the entire continuous hyperparameter space rather than a limited, predefined grid. However, this approach is challenging in practice for deep neural networks because computing the evidence involves a high-dimensional integral that is generally intractable for models with millions of parameters.

To make this feasible, the method leverages variational Bayesian techniques, specifically using tractable lower bounds on the evidence, to enable efficient hyperparameter selection during fine-tuning of deep neural image classifiers [29, 30]. Experimental results indicate that this approach can significantly reduce training time compared to traditional grid search-based methods—reducing total training time from many hours to under three in some cases—while maintaining or improving accuracy, especially when labeled data is limited. This makes Bayesian evidence maximization a practical and cost-effective solution for hyperparameter tuning in transfer learning scenarios.

4.2.2 Methodology

Variational methods for posterior estimation

For complex probabilistic models, directly evaluating the posterior distribution $p(y_{1:N}, w, V)$ is computationally infeasible because it requires calculating the normalization constant, or evidence, $p(w, V|y_{1:N})$, which itself is a high-dimensional integral over all possible parameter values. This intractability makes both evaluating the posterior for specific parameter values and sampling from it extremely challenging. To address this, variational inference (VI) provides a

Table 4.7: Possible priors.

Method	$p(w)$	Init.
L2-zero	$\mathcal{N}(0, \lambda I)$	μ
L2-SP	$\mathcal{N}(\mu, \lambda I)$	μ
PTYL	$\mathcal{N}(\mu, \lambda \Sigma)$	μ

practical alternative by approximating the true posterior with a simpler, tractable distribution $q(w, V) \approx p(w, V | y_{1:N})$ that is chosen from a parameterized family. In this work, we represent the variational posterior q as a product of two independent Gaussian distributions, $q(w)$ and $q(V)$. Specifically, we define

$$q(w, V) = q(w)q(V), q(w) = \mathcal{N}(w | \bar{w}, \bar{\sigma}^2 I), q(V) = \mathcal{N}(\text{vec}(V) | \text{vec}(\bar{V}), \bar{\sigma}^2 I). \quad (4.6)$$

This factorized Gaussian structure is not only computationally convenient, but—as we will demonstrate later—enables several intermediate computations to be carried out analytically in closed form. We denote the free parameters of q with bars to distinguish them from model parameters. Weights $\bar{w}, \bar{V}$ represent means of the backbone and classifier head. Scalar $\bar{\sigma}^2 > 0$ controls variance around these means. We use one variance parameter for simplicity; future work could allow separate parameters for different classifier parts.

The goal is to make $q(w, V)$ as close as possible to the true posterior, typically by minimizing the Kullback–Leibler (KL) divergence between them. For two general distributions $p(x)$ and $q(x)$, KL divergence is defined as

$$\text{KL}(p \| q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx \quad (4.7)$$

The KL divergence term between the true posterior and the variational posterior cannot be computed analytically because the true posterior $p(w, V | y_{1:N})$ is intractable. However, it can be shown that minimizing the KL divergence between the variational posterior $q(w, V)$ and the true posterior $p(w, V | y_{1:N})$ is equivalent to maximizing the so-called evidence lower bound (ELBo). The ELBo is an objective function that trades off how well the variational posterior explains the observed data and how closely it matches the prior [29]. The ELBo is defined as

$$J_{\text{ELBo}} := \mathbb{E}_{q(w,V)} \left[\sum_{i=1}^N \log p(y_i|w, V) \right] - \mathbb{KL}(q(w)||p(w|\lambda)) - \mathbb{KL}(q(V)||p(V|\tau)). \quad (4.8)$$

This objective is a function of data $y_{1:N}, x_{1:N}$, variational posterior parameters $\bar{w}, \bar{V}, \bar{\sigma}$, and prior hyperparameters λ, τ . Maximizing the ELBo thus identifies the best approximation within the chosen family, and it is mathematically guaranteed to be a lower bound on the true log-evidence of the data. This property not only guides the optimization but also enables principled model and hyperparameter selection, with theoretical guarantees even when the model is misspecified.

When optimizing the ELBo for neural network classifiers, the two KL divergence terms are straightforward to compute and differentiate because they involve KL divergences between multivariate Gaussian distributions, which have closed-form solutions. The primary computational challenge lies in the expected log likelihood term, which, due to the non-conjugacy introduced by neural networks, cannot be evaluated analytically. To address this, the reparameterization trick is used [31, 32], as suggested in foundational works on variational inference for deep models [33]. To implement this estimator for this likelihood term, we first draw vectors the same size as w and $\text{vec}(V)$, denoted ϵ^w and ϵ^V , from a standard normal. Then, we transform to samples of w, V from q via well-known transformations (e.g., for the backbone: $w \leftarrow \bar{w} + \bar{\sigma}\epsilon^w$). Averaging over such independent and identically distributed (IID) samples w, V from q allows a Monte Carlo approximation of the expected log likelihood. Gradients follow via automatic differentiation of that Monte Carlo estimator’s computational graph. Empirically, using *just one sample* per training step is sufficient and most efficient.

Learning key hyperparameters λ, τ with the ELBo

While most work on variational inference focuses on learning only the posterior over model parameters, it is also straightforward to optimize the ELBo objective with respect to the prior hyperparameters λ and τ . Recall that the term $-\mathbb{KL}(q(w)||p(w))$ remains

tractable in this setting, since both the approximate posterior and the prior are multivariate Gaussians.

For instance, in our particular model in Eq. (4.2), the KL divergence between two Gaussians [34] simplifies for the backbone KL term as:

$$-\mathbb{KL}(q(w)||p(w)) = -\frac{1}{2} \left[\frac{\bar{\sigma}^2}{\lambda} \text{Tr}(\Sigma_p^{-1}) + \frac{1}{\lambda} (\mu_p - \bar{w})^T \Sigma_p^{-1} (\mu_p - \bar{w}) - D + \log \left(\frac{\lambda^D \det(\Sigma_p)}{\bar{\sigma}^{2D}} \right) \right]. \quad (4.9)$$

A simpler expression is possible for the KL over the classifier head weights V as well:

$$-\mathbb{KL}(q(V)||p(V)) = -\frac{1}{2} \left[\frac{\bar{\sigma}^2}{\tau} D + \frac{1}{\tau} \|\text{vec}(\bar{V})\|_2^2 - D + \log \left(\frac{\tau^D}{\bar{\sigma}^{2D}} \right) \right]. \quad (4.10)$$

Closed-form updates. To find an optimal λ value with respect to the J_{ELBo} , notice that of the 3 additive terms in Eq. (4.8), only the KL term between $q(w)$ and $p(w)$ involves λ . We solve for λ by taking the gradient of the KL term with respect to λ , setting to zero, and solving, with assurances of a local maximum of J_{ELBo} via a second derivative test (Appendix A.1) [35, 36]. The gradient is

$$\nabla_{\lambda} -\mathbb{KL}(q(w)||p(w)) = -\frac{1}{2} \left[-\frac{\bar{\sigma}^2}{\lambda^2} \text{Tr}(\Sigma_p^{-1}) - \frac{1}{\lambda^2} (\mu_p - \bar{w})^T \Sigma_p^{-1} (\mu_p - \bar{w}) + \frac{D}{\lambda} \right]. \quad (4.11)$$

Setting $\nabla_{\lambda} -\mathbb{KL}(q(w)||p(w)) = 0$ and solving for λ , we get

$$\lambda^* = \frac{1}{D} \left[\bar{\sigma}^2 \text{Tr}(\Sigma_p^{-1}) + (\mu_p - \bar{w})^T \Sigma_p^{-1} (\mu_p - \bar{w}) \right]. \quad (4.12)$$

Similar updates can be derived for τ :

$$\nabla_{\tau} -\mathbb{KL}(q(V)||p(V)) = -\frac{1}{2} \left[-\frac{\bar{\sigma}^2}{\tau^2} D - \frac{1}{\tau^2} \|\text{vec}(\bar{V})\|_2^2 + \frac{D}{\tau} \right]. \quad (4.13)$$

Setting $\nabla_{\tau} - \mathbb{KL}(q(V)||p(V)) = 0$ and solving for τ , we get

$$\tau^* = \bar{\sigma}^2 + \frac{1}{D} \|\text{vec}(\bar{V})\|_2^2. \quad (4.14)$$

Inflating dataset size to select complex models with better generalization

In transfer learning scenarios where only a few dozen or hundred labeled examples are available, classical statistical learning theory would typically not recommend using deep neural networks with millions of parameters, as the number of parameters far exceeds the number of data points (when $D \gg N$). Traditionally, this would risk severe overfitting and poor generalization. However, recent trends in deep learning have shown that, with careful fine-tuning, large pre-trained models can achieve strong generalization even in these data-scarce settings [27, 37].

When ELBo is maximized naively in the overparameterized regime, the objective becomes overly conservative. Specifically, the KL divergence term, which penalizes deviation from the prior, can dominate the objective, leading to the selection of overly simple models that perform poorly on the target task. This is because, with so few data points, the likelihood term in Eq. (4.8) is too weak to counterbalance the strong regularization imposed by the KL term, resulting in underfitting.

To address this, we modified the standard ELBo objective by introducing a scaling factor $\kappa \geq 1$ on the likelihood term.

$$J_{\text{DE ELBo}} := \kappa \mathbb{E}_{q(w,V)} \left[\sum_{i=1}^N \log p(y_i|w, V) \right] - \mathbb{KL}(q(w, V)||p(w, V)) \quad (4.15)$$

This adjustment increases the influence of the data fit relative to the regularization, making it possible to select more complex models that better capture the structure in the limited available data (Fig. 4.3). In practice, when D is in the millions and N is in the tens or hundreds, setting proportional to can restore a more appropriate balance between fitting the data and regularizing the model, leading to improved performance on the target classification task. We call this objective the *data-emphasized ELBo* (DE ELBo).

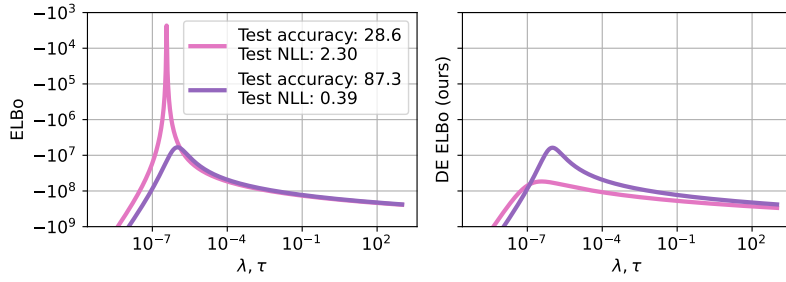


Figure 4.3: Model selection comparison between the ELBo (left) and our *data-emphasized ELBo* (DE ELBo) (right) for two ResNet-50s trained on CIFAR-10 $N = 1000$. For both models, we fix the estimated posterior q and vary λ, τ .

We use it with $\kappa = D/N$ for all training steps, as well as ultimate hyperparameter/model selection.

In Fig. 4.3, we compare the cases of $\kappa = 1$ (left panel) and $\kappa = D/N$ (right) on CIFAR-10 for L2-SP. In each plot, we compare two possible q , “A” (in pink) and “B” (in purple), across a range of λ, τ values. Version A sets $\bar{w}_A, \bar{V}_A, \bar{\sigma}_A$ to a solution favored by ELBo, with known test-set accuracy 28.6%. Version B sets $\bar{w}_B, \bar{V}_B, \bar{\sigma}_B$ to values favored by our (DE ELBo), with known test-set accuracy of 87.3%. We see the poor-accuracy version A model is strongly favored by the conventional ELBo, while the more accurate version B model is favored by our recommended DE ELBo with $\kappa = D/N$.

Overall algorithm and implementation

The fitting algorithm in this work uses stochastic gradient descent (SGD) to optimize the parameters of the variational posterior, specifically the means $\bar{w}$, $\bar{V}$, and the standard deviation parameterized as $\bar{\sigma} = \log(1 + \exp(\rho))$ with free parameter $\rho \in \mathbb{R}$, which ensures $\bar{\sigma}(\rho)$ is always positive during gradient descent. Before each SGD update to these variational parameters, closed-form updates are performed for the hyperparameters λ and τ , governed by Eq. (4.12) and Eq. (4.14) respectively, and this step is independent of the scaling factor κ . The optimization process continues until a predetermined maximum number of iterations is reached.

This approach enables the method to produce both an estimated posterior q and learned hyperparameters λ and τ in a single run of gradient descent. For prediction, the posterior means $\bar{w}$ and $\bar{V}$ are used. The method is compared to a baseline that performs

MAP point estimation for w and V , requiring separate SGD runs for each configuration of λ and τ on a fixed grid—a common strategy in recent transfer learning research [5, 38]. Both methods use the ResNet-50 architecture as the backbone network [16].

The performance of both the proposed method and the baseline depends on the choice of learning rate, so each run searches over four candidate learning rates, selecting the best one based on either the modified ELBo (for the proposed method) or validation-set likelihood (for the baseline). Both approaches are compatible with any of three backbone prior settings described in Table 4.7. For the PTYL method [5], the implementation uses corrected code [38] to ensure proper scaling of the learned low-rank covariance matrix. Efficient computation of key quantities for the KL divergence term, such as the trace of the inverse, squared Mahalanobis distance, and log determinant of the low-rank covariance matrix, is achieved using the Woodbury matrix identity, trace properties, and the matrix determinant lemma. Further implementation details are provided in Appendix A.2.

Model	Method	Avg. SGD runtime	LR Search	λ, τ Search	Total GS time
L2-SP	MAP + GS	39 mins. 11 secs.	4	6	16 hrs. 15 mins.
	DE ELBo	39 mins. 0 secs.	4	n/a	2 hrs. 36 mins.
PTYL	MAP + GS	37 mins. 15 secs.	4	60	149 hrs. 36 mins.
	DE ELBo	40 mins. 0 secs.	4	n/a	2 hrs. 39 mins.

Table 4.8: Computational time comparison between methods for transfer learning with informative priors. “Search” indicates the dimension of the search space.

4.2.3 Results

The result of this work can be summarized in three main points.

First, the DE ELBo approach enables effective learning of regularization strength while maintaining high task accuracy, especially in settings where the number of model parameters greatly exceeds the number of training examples. In such overparameterized regimes, the standard ELBo tends to favor overly simple models due to the strong influence of the prior, which can hurt performance on the target task (Fig. 4.3). By scaling the likelihood term with a factor equal to the ratio of parameter dimension to data points, this method ensures that the data fit receives appropriate emphasis compared to

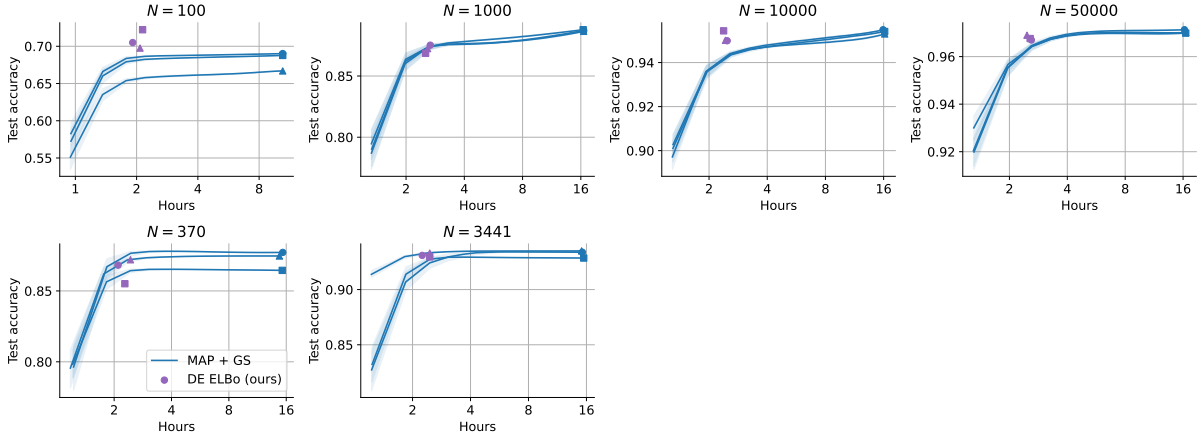


Figure 4.4: Test-set accuracy on CIFAR-10 (top row) and Oxford-IIIT Pet (bottom row) over training time for L2-SP with *MAP + grid search (GS)* and our *data-emphasized ELBo (DE ELBo)*. We run each method on 3 separate training sets of size N (3 different marker styles).

the regularization term. As a result, the modified objective can jointly learn both the backbone and classifier head weights $\bar{w}, \bar{V}$, as well as important hyperparameters λ, τ , in a way that leads to strong generalization and effectiveness on downstream tasks.

Second, DE ELBo offers a significant computational advantage over traditional methods that rely on grid search for hyperparameter tuning (Table 4.8). Specifically, a single SGD run using DE ELBo requires roughly the same computational resources as an SGD run for standard MAP estimation. However, when hyperparameters such as λ and τ need to be selected for the MAP baseline, grid search becomes extremely costly in terms of runtime. For instance, in the PTYL setting, grid search can take over 149 hours to complete, whereas the DE ELBo method achieves comparable results in less than 3 hours. This efficiency is achieved without compromising performance, making DE ELBo a practical and scalable alternative to exhaustive grid search.

Third, DE ELBo approach achieves test accuracy on par with established methods but requires significantly less compute time (Fig. 4.4). When evaluated on the CIFAR-10 [10] and Oxford-IIIT Pet datasets [12], the method delivers heldout accuracy that matches the results of grid search-based baselines, yet completes training much faster. This efficiency is particularly notable when the amount of labeled training data is small; for example, on CIFAR-10 with only 100 training examples, DE ELBo can even outperform the grid search baseline. Final test accuracies reported in Table 4.9 and Table 4.10 confirm that

this approach provides competitive performance without the computational burden of exhaustive grid search.

Model	Method	$N = 100$ (10/cl.)	1000 (100/cl.)	10000 (1k/cl.)	50000 (5k/cl.)
L2-zero	MAP + GS	67.7 (66.0-68.6)	87.8 (87.5-88.4)	95.0 (94.4-95.5)	97.2 (97.1-97.2)
	DE ELBo	60.9 (58.9-63.1)	87.2 (87.0-87.4)	91.2 (90.7-92.0)	93.2 (93.0-93.3)
L2-SP	MAP + GS	68.1 (66.7-68.9)	87.3 (87.2-87.3)	95.3 (95.1-95.7)	97.1 (97.0-97.1)
	DE ELBo	70.6 (68.7-72.7)	87.2 (86.8-87.4)	95.0 (94.8-95.2)	96.8 (96.7-96.9)
PTYL	MAP + GS	67.5 (65.7-68.4)	87.9 (86.9-89.2)	95.2 (95.0-95.4)	97.3 (97.3-97.3)
	DE ELBo	70.6 (68.7-72.6)	87.2 (86.9-87.6)	95.1 (94.9-95.4)	96.9 (96.8-96.9)

Table 4.9: Accuracy on CIFAR-10 test set for different probabilistic models and methods.

Model	Method	$N = 370$ (10/cl.)	3441(93/cl.)
L2-zero	MAP + GS	87.2 (86.6-87.5)	93.2 (93.0-93.4)
	DE ELBo	81.8 (80.3-83.6)	92.6 (92.1-92.4)
L2-SP	MAP + GS	87.2 (86.6-87.6)	93.2 (93.0-93.4)
	DE ELBo	86.6 (85.7-87.3)	93.2 (93.0-93.3)
PTYL	MAP + GS	87.1 (86.6-87.5)	92.2 (90.7-93.2)
	DE ELBo	86.6 (85.7-87.3)	93.2 (93.0-93.3)

Table 4.10: Accuracy on Oxford-IIIT Pet test set for different probabilistic models and methods.

4.2.4 Conclusion

We propose a novel alternative to grid search for selecting regularization hyperparameters in deep neural networks, by introducing DE ELBo that increases the weight of the data likelihood term relative to the prior, which is particularly beneficial when training data is limited or the model is highly overparameterized. This adjustment improves model selection, enabling the approach to achieve heldout accuracy on benchmarks like CIFAR-10 and Oxford-IIIT Pet that matches existing methods, but with significantly reduced compute time-cutting grid search runtimes from over 88 hours to under 3 hours.

Bibliography

- [1] L. Brigato and L. Iocchi. “A Close Look at Deep Learning with Small Data”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021.
- [2] J. Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [3] F. Zhuang et al. “A Comprehensive Survey on Transfer Learning”. *Proceedings of the IEEE* 109 (2020), pp. 43–76.
- [4] K. S. Murphy. “Probabilistic Machine Learning: An Introduction”. In: MIT Press, 2022. Chap. 19.2 Transfer learning.
- [5] R. Shwartz-Ziv et al. “Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.
- [6] T. Chen, E. Fox, and C. Guestrin. “Stochastic Gradient Hamiltonian Monte Carlo”. In: *International Conference on Machine Learning (ICML)*. 2014.
- [7] J. Pineau et al. “Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)”. *The Journal of Machine Learning Research* 22 (2021), pp. 7459–7478.
- [8] K. P. Murphy. “Machine Learning: A Probabilistic Perspective”. In: MIT Press, 2012. Chap. 3.3.3 Posterior.
- [9] C. Chelba and A. Acero. “Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot”. *Computer Speech & Language* 20 (2006), pp. 382–399.
- [10] A. Krizhevsky, V. Nair, and G. Hinton. “CIFAR-10 (Canadian Institute for Advanced Research)” (2010).
- [11] M.-E. Nilsback and A. Zisserman. “Automated Flower Classification over a Large Number of Classes”. In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*. 2008.
- [12] O. M. Parkhi et al. “Cats And Dogs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.
- [13] S. Maji et al. “Fine-Grained Visual Classification of Aircraft”. *arXiv preprint arXiv:1306.5151* (2013).
- [14] P. Tschandl, C. Rosendahl, and H. Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. *Scientific Data* 5 (2018), pp. 1–9.
- [15] N. Codella et al. “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”. *arXiv preprint arXiv:1902.03368* (2019).
- [16] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [17] I. Loshchilov and F. Hutter. “SGDR: Stochastic Gradient Descent with Warm Restarts”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2016.

-
- [18] G. Kaplun et al. “SubTuning: Efficient Finetuning for Multi-Task Learning”. *arXiv preprint arXiv:2302.06354* (2023).
- [19] E. Harvey, M. Petrov, and M. C. Hughes. “Transfer Learning with Informative Priors: Simple Baselines Better than Previously Reported”. *Transactions on Machine Learning Research* ().
- [20] K. S. Murphy. “Probabilistic Machine Learning: An Introduction”. In: MIT Press, 2022. Chap. 4.5 Regularization.
- [21] T. Hastie, R. Tibshirani, and J. Friedman. “The Elements of Statistical Learning”. In: Second. Springer Series in Statistics. Springer, 2009. Chap. Sec. 3.4 Shrinkage Methods.
- [22] D. Kobak, J. Lomond, and B. Sanchez. “The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative due to the Implicit Ridge Regularization”. *Journal of Machine Learning Research (JMLR)* 21 (2020), pp. 1–16.
- [23] A. Krogh and J. A. Hertz. “A Simple Weight Decay Can Improve Generalization”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 1991.
- [24] I. Goodfellow, Y. Bengio, and A. Courville. “Deep Learning”. In: MIT Press, 2016. Chap. 7 Regularization for Deep Learning.
- [25] S. Raschka. “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning”. *arXiv preprint arXiv:1811.12808* (2018).
- [26] K. S. Murphy. “Probabilistic Machine Learning: An Introduction”. In: MIT Press, 2022. Chap. 4.5.4 Picking the regularizer using a validation set.
- [27] L. Xuhong, Y. Grandvalet, and F. Davoine. “Explicit Inductive Bias for Transfer Learning with Convolutional Networks”. In: *International Conference on Machine Learning (ICML)*. 2018.
- [28] S. Lotfi et al. “Bayesian Model Selection, the Marginal Likelihood, and Generalization”. In: *International Conference on Machine Learning (ICML)*. 2022.
- [29] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. “Variational Inference: A Review for Statisticians”. *Journal of the American Statistical Association* 112 (2017), pp. 859–877.
- [30] M. I. Jordan et al. “An introduction to variational methods for graphical models”. *Machine learning* 37 (1999), pp. 183–233.
- [31] M. Xu et al. “Variance Reduction Properties of the Reparameterization Trick”. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2019.
- [32] S. Mohamed et al. “Monte Carlo Gradient Estimation in Machine Learning”. *Journal of Machine Learning Research (JMLR)* 21 (2020).
- [33] C. Blundell et al. “Weight Uncertainty in Neural Networks”. In: *International Conference on Machine Learning (ICML)*. 2015.
- [34] K. S. Murphy. “Probabilistic Machine Learning: An Introduction”. In: MIT Press, 2022. Chap. 6.2.3 Example: KL divergence between two Gaussians.

- [35] E. Harvey, M. Petrov, and M. C. Hughes. “Learning the Regularization Strength for Deep Fine-Tuning via a Data-Emphasized Variational Objective”. In: *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.
- [36] E. Harvey, M. Petrov, and M. C. Hughes. “Learning Hyperparameters via a Data-Emphasized Variational Objective”. *arXiv preprint arXiv:2502.01861* (2025).
- [37] A. Sharif Razavian et al. “CNN Features off-the-shelf: an Astounding Baseline for Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2014.
- [38] E. Harvey, M. Petrov, and M. C. Hughes. “Transfer Learning with Informative Priors: Simple Baselines Better than Previously Reported”. *Transactions on Machine Learning Research (TMLR)* (2024). Reproducibility Certification.

Appendix A

Mathematical details

A.1 Second Derivative Test

The second derivative is

$$\nabla_{\lambda}^2 - \mathbb{KL}(q(w)||p(w)) = -\frac{1}{2} \left[\frac{2\bar{\sigma}^2}{\lambda^3} \text{Tr}(\Sigma_p^{-1}) + \frac{2}{\lambda^3} (\mu_p - \bar{w})^T \Sigma_p^{-1} (\mu_p - \bar{w}) - \frac{D}{\lambda^2} \right] \quad (\text{A.1})$$

$$= -\frac{1}{2} \left[\frac{2D}{\lambda^3} \frac{1}{D} (\bar{\sigma}^2 \text{Tr}(\Sigma_p^{-1}) + (\mu_p - \bar{w})^T \Sigma_p^{-1} (\mu_p - \bar{w})) - \frac{D}{\lambda^2} \right] \quad (\text{A.2})$$

$$= -\frac{1}{2} \left[\frac{2D}{\lambda^3} \lambda^* - \frac{D}{\lambda^2} \right]. \quad (\text{A.3})$$

Plugging in λ^* and simplifying, we get

$$\nabla_{\lambda}^2 - \mathbb{KL}(q(w)||p(w|\lambda^*)) = -\frac{D}{2} \frac{1}{\lambda^{*2}} \quad (\text{A.4})$$

This expression is always negative, indicating that λ^* is a local maximum of J_{ELBo} .

The second derivative is

$$\nabla_{\tau}^2 - \mathbb{KL}(q(V)||p(V)) = -\frac{1}{2} \left[\frac{2\bar{\sigma}^2}{\tau^3} D + \frac{2}{\tau^3} \|\text{vec}(\bar{V})\|_2^2 - \frac{D}{\tau^2} \right] \quad (\text{A.5})$$

$$= -\frac{1}{2} \left[\frac{2D}{\tau^3} \left(\bar{\sigma}^2 + \frac{1}{D} \|\text{vec}(\bar{V})\|_2^2 \right) - \frac{D}{\tau^2} \right] \quad (\text{A.6})$$

$$= -\frac{1}{2} \left[\frac{2D}{\tau^3} \tau^* - \frac{D}{\tau^2} \right]. \quad (\text{A.7})$$

Plugging in τ^* and simplifying, we get

$$\nabla_{\tau}^2 - \mathbb{KL}(q(V)||p(V|\tau^*)) = -\frac{D}{2} \frac{1}{\tau^{*2}}. \quad (\text{A.8})$$

This expression is always negative, indicating that τ^* is a local maximum of J_{ELBo} .

A.2 Low-Rank Σ_p

The PTYL method [1] uses Stochastic Weight Averaging-Gaussian (SWAG) [2] to approximate the posterior distribution $p(w|\mathcal{D}_S)$ of the source data $\mathcal{D}_S$ with a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ where μ is the learned mean and $\Sigma = \frac{1}{2}(\Sigma_{\text{diag}} + \Sigma_{\text{LR}})$ is a representation of a covariance matrix with both diagonal and *low-rank* components. The LR covariance has the form $\Sigma_{\text{LR}} = \frac{1}{K-1} Q Q^T$, where $Q \in \mathbb{R}^{D \times K}$.

We use the Woodbury matrix identity [3], trace properties, and the matrix determinant lemma to compute the trace of the inverse, squared Mahalanobis distance, and log determinant of the low-rank covariance matrix for the KL term.

The trace and log determinant of the low-rank covariance matrix can be calculated once and used during training. Just like in PTYL method, the squared Mahalanobis distance needs to be re-evaluated every iteration of gradient descent.

A.2.1 Trace of the inverse

We compute the trace of the inverse of the low-rank covariance matrix using the Woodbury matrix identity and trace properties.

$$\begin{aligned}
\text{Tr}(\Sigma_p^{-1}) &= \text{Tr}((A + UCV)^{-1}) \\
&= \text{Tr}(A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}) && \text{Woodbury matrix identity} \\
&= \text{Tr}(A^{-1}) - \text{Tr}(A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}) && \text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B) \\
&= \text{Tr}(A^{-1}) - \text{Tr}((C^{-1} + VA^{-1}U)^{-1}VA^{-1}A^{-1}U) && \text{Tr}(AB) = \text{Tr}(BA)
\end{aligned}$$

where $A = \frac{1}{2}\Sigma_{\text{diag}}$, $C = I_K$, $U = \frac{1}{\sqrt{2K-2}}Q$, and $V = \frac{1}{\sqrt{2K-2}}Q^T$. The last trace property, lets us compute the trace of the inverse of the low-rank covariance matrix without having to store a $D \times D$ covariance matrix.

A.2.2 Squared Mahalanobis distance

We compute the squared Mahalanobis distance $(\mu_p - \bar{w})^T \Sigma_p^{-1} (\mu_p - \bar{w})$ by distributing the mean difference vector into the Woodbury matrix identity.

$$\begin{aligned}
\Sigma_p^{-1} &= (A + UCV)^{-1} \\
&= (A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}) && \text{Woodbury matrix identity}
\end{aligned}$$

A.2.3 Log determinant

We compute the log determinant of the low-rank covariance matrix using the matrix determinant lemma.

$$\begin{aligned}
\log \det(\Sigma_p) &= \log \det(A + UV) \\
&= \log(\det(I_K + VA^{-1}U) \det(A)) && \text{Matrix determinant lemma}
\end{aligned}$$

Appendix B

Classification details

B.1 Dataset Details

Our experiments include a replication of CIFAR-10 [4] experiments from [1], similar Oxford Flowers [5] experiments to [1], similar Oxford-IIIT Pets [6] experiments to [1], additional FGVC-Aircraft [7] experiments, and additional HAM10000 [8, 9] experiments.

For Oxford Flowers experiments, we modify prior experiments to enforce classes are uniformly distributed in the training data, that at least one sample from each class is included in all tested values of n , and that the validation set is proportional with the training set. [1] validate on the predefined validation set specified by the dataset creators. In our minds, it is not realistic to train on a range of dataset sizes of 102 or 510 images but validate on 1020 images. Practitioners looking for guidance in how to use transfer learning on their dataset will be unlikely to put more images in validation than training. Instead, we ignore the predefined validation set and use 1/5 of the training set for our two-phase training pipeline.

For Oxford-IIIT Pets experiments, we modify prior experiments to enforce classes are uniformly distributed in the training data, and that at least one sample from each class is included in all tested values of n .

For FGVC-Aircraft experiments, we combine the training and validation sets specified by the creators to enforce that the validation set is proportional to the training set.

For HAM10000 experiments, we form a classification task using classes with sufficient data (at least 10% prevalence). One additional “other” class was formed from all other classes that did not have sufficient prevalence. For the validation set, we ensure each patient’s data belongs to exactly one split and stratify by class to ensure comparable class frequencies.

Table B.1: Distribution of labels in HAM10000 training set.

Label	Percent of training set
Melanocytic	66.95%
Melanoma	11.11%
Benign keratosis	10.97%
Other	10.97%

We use the same preprocessing steps for all five datasets. For each distinct training set size n , we compute the mean and standard deviation of each channel to normalize images. During fine-tuning we resize the images to 256x256 pixels, perform random cropping to 224x224, and perform horizontal flips. At test time, we resize the images to 256x256 pixels and center crop to 224x224.

B.2 Classifier Details

We use SGD with a Nesterov momentum parameter of 0.9 for optimization. We train for 6,000 steps using a cosine annealing learning rate [10]. We select the initial learning rate from 4 logarithmically spaced values between 0.1 to 0.0001 and the weight decay from 5 logarithmically spaced values between $1e^{-2}$ to $1e^{-6}$, as well as without weight decay. For *LearnedPriorLR*, we select the covariance scaling factor from 10 logarithmically spaced values between $1e^0$ to $1e^9$.

Code fixes. The released code from `citeshwartz2022pre` performs inconsistent scaling by λ of the low-rank and diagonal components of the covariance matrix of their informed prior. Looking at their code (see line 40 of `utils.py`), they obtain Σ_{LR} by rescaling the low-rank matrix Q by λ before computing QQ^T , which yields an overall covariance matrix $\frac{\lambda}{2}\Sigma_{\text{diag}} + \frac{\lambda^2}{2}\Sigma_{\text{LR}}$. To remedy this, we rescale the low-rank matrix Q by $\sqrt{\lambda}$ instead of λ ,

obtaining the intended $\frac{\lambda}{2}\Sigma_{\text{diag}} + \frac{\lambda}{2}\Sigma_{\text{LR}}$. We also remove the redundant “double prior” due to their implementation’s use of weight decay on weights w in addition to the informative prior.

Varying train set size n . For each dataset, we setup a range of transfer learning tasks by artificially varying n , the number of available samples for training. In all cases, we ensure *balanced class frequencies*, ensuring the same number of images per class. For CIFAR-10, by selecting $n \in \{10, 100, 1000, 10000, 50000\}$, we cover exactly 1, 10, 100, 1000, and all available examples per class. For Oxford Flowers, we select $n \in \{102, 510, 1020\}$ to cover 1, 5, and 10 examples per class. For Oxford-IIIT Pets, we select $n \in \{37, 370, 3441\}$ to cover 1, 10, and 93 examples per class (93 is the maximum possible value that keeps exact balance). For FGCV-Aircraft, we select $n \in \{100, 500, 1000, 5000\}$ to cover 1, 5, 10, and 50 examples per class. For HAM10000, we select $n \in \{100, 1000\}$. We emphasize that the 10-1000 examples per class regime is likely of most practical interest for many applications.

Repeated trials. At each chosen value of n , we build 3 separate “replicate” training sets by randomly sampling n image-label pairs (without replacement) from the predefined complete training set. We then separately perform all training (including hyperparameter tuning) on each replicate. We emphasize we are careful to ensure similar class-composition for all replicates.

While tuning hyperparameters, we hold out 1/5 of the training set for validation, ensuring balanced class frequencies between sets. At the smallest tested values of n suggested by [1], the number of training samples is equal to the number of classes, and thus any images assigned to the validation set will have their corresponding class not represented in the training data.

After selecting the optimal hyperparameters from the validation set NLL, we retrain the model using the selected hyperparameters on the combined set of all n images (merging training and validation). All results report performance on the task in question’s predefined test set.

StdPrior. Stochastic gradient descent (SGD) in PyTorch [11] implements weight decay via an efficient direct modification of the gradient vector. Given decay parameter $\alpha > 0$, at iteration t the optimizer internally performs an additive incremental update of the gradient with respect to backbone weights: $\nabla_w \ell(w_{t-1}) \leftarrow \nabla_w \ell(w_{t-1}) + \alpha w_{t-1}$ where $\ell(w)$ is the (unpenalized) loss function (e.g. cross entropy). The gradient with respect to classifier parameters V is updated similarly. This direct modification is equivalent to, but more efficient than, automatic differentiation of a penalized loss that includes an L2 penalty term for each of w and V :

$$L(w, V) := \underbrace{-\frac{1}{n} \sum_{i=1}^n \log p(y_i | x_i, w, V)}_{\ell(w, V)} + \frac{\alpha}{2} \|w\|^2 + \frac{\alpha}{2} \|\text{vec}(V)\|^2. \quad (\text{B.1})$$

Weight decay in PyTorch has a Bayesian interpretation as maximum a-posteriori (MAP) estimation of parameters w, V , under the assumption that the weights w, V each follow a Gaussian prior $\mathcal{N}(0, \tau I)$ with scalar precision parameter $\tau = \frac{1}{n\alpha}$. Rescaling by n ensures that minimizing the above loss is equivalent to maximizing the per-example MAP objective: $\frac{1}{n} [\log p(y|x, w, V) + \log p(w, V)]$.

We use SimCLR Resnet-50 pre-trained initialization from [1] to initialize weights w .

LearnedPriorIso. For *LearnedPriorIso*, we minimize the following MAP objective

$$L(w, V) := -\frac{1}{n} \sum_{i=1}^n \log p(y_i | x_i, w, V) + \frac{\alpha}{2} \|w - \mu\|^2 + \frac{\alpha}{2} \|\text{vec}(V)\|^2 \quad (\text{B.2})$$

so that the weight decay search space is the same as SGD in PyTorch. We use SimCLR Resnet-50 pre-trained initialization from [1] as the mean μ and to initialize weights w .

LearnedPriorLR. For *LearnedPriorLR*, we minimize the following MAP objective

$$L(w, V) := -\frac{1}{n} \sum_{i=1}^n \log p(y_i | x_i, w, V) - \frac{1}{n} \log \mathcal{N}(w | \mu, \lambda \Sigma) + \frac{\alpha}{2} \|\text{vec}(V)\|^2. \quad (\text{B.3})$$

Although not discussed in [1], they add ε to the prior variance (they set the default value of ε to 0.1). We include the addition of ε to the prior variance by adding $\frac{\lambda}{2}\Sigma_{\text{diag}} + \varepsilon I$. We note that adding ε to the prior variance has a significant effect if the covariance scaling factor is small (e.g., $\lambda = 1$).

Bibliography

- [1] R. Shwartz-Ziv et al. “Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.
- [2] W. J. Maddox et al. “A Simple Baseline for Bayesian Uncertainty in Deep Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.
- [3] M. A. Woodbury. *Inverting Modified Matrices*. Department of Statistics, Princeton University, 1950.
- [4] A. Krizhevsky, V. Nair, and G. Hinton. “CIFAR-10 (Canadian Institute for Advanced Research)” (2010).
- [5] M.-E. Nilsback and A. Zisserman. “Automated Flower Classification over a Large Number of Classes”. In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*. 2008.
- [6] O. M. Parkhi et al. “Cats And Dogs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.
- [7] S. Maji et al. “Fine-Grained Visual Classification of Aircraft”. *arXiv preprint arXiv:1306.5151* (2013).
- [8] P. Tschandl, C. Rosendahl, and H. Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. *Scientific Data* 5 (2018), pp. 1–9.
- [9] N. Codella et al. “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”. *arXiv preprint arXiv:1902.03368* (2019).
- [10] I. Loshchilov and F. Hutter. “SGDR: Stochastic Gradient Descent with Warm Restarts”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2016.
- [11] A. Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

Funding Disclosure

The works presented in this thesis were supported by the following grants:

- NIH R01 CA262147
- NSF CMMI 1937373
- NSF CMMI 2224708
- MLSC women's health program

Publication list

- Detection of Human Bladder Epithelial Cancerous Cells with Atomic Force Microscopy and Machine Learning.

Mikhail Petrov, Nadezhda Makarova, Amir Monemian, Jean Pham, Małgorzata Lekka, and Igor Sokolov *Cells*, 14, 14 - December 2024.

- Mechanical spectroscopy of materials using atomic force microscopy (AFM-MS).

Mikhail Petrov, Daniel Canena , Nikita Kulachenkov, Nishant Kumar, Pierre Nickmilder, Philippe Leclère, Igor Sokolov *Materials today*, 80, 218-225 - September 2024.

- Transfer Learning with Informative Priors: Simple Baselines Better than Previously Reported.

Ethan Harvey*, **Mikhail Petrov*** and Michael C Hughes. *Transactions on Machine Learning Research*, 2835-8856 - May 2024.

- Machine Learning Allows for Distinguishing Precancerous and Cancerous Human Epithelial Cervical Cells Using High-Resolution AFM Imaging of Adhesion Maps.

Mikhail Petrov and Igor Sokolov. *Cells*, 12(21), 2536 - October 2023.

- Identification of Geometrical Features of Cell Surface Responsible for Cancer Aggressiveness: Machine Learning Analysis of Atomic Force Microscopy Images of Human Colorectal Epithelial Cells.

Mikhail Petrov and Igor Sokolov. *Biomedicines* 11(1) - January 2023.

Conference Presentations

- **Talk:** “Identification of geometrical features of cell surface responsible for cancer aggressiveness: Machine learning analysis of atomic force microscopy images”, at the *2024 MRS Fall Meeting & Exhibit*, December 1-6 2024 in Boston, MA, USA.
- **Poster:** “Transfer Learning with Informative Priors: Simple Baselines Better than Previously Reported”, at the *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, December 10-15 2024 in Vancouver, BC, Canada.
- **Poster:** “Learning the Regularization Strength for Deep Fine-Tuning via a Data-Emphasized Variational Objective”, at the *Fine-Tuning in Modern Machine Learning: Principles and Scalability (FITML)*, December 14 2024 in Vancouver, BC, Canada.
- **Talk:** “Recognition of Polymers with Nanometer Resolution: Mechano-Spectroscopy of Polymers with AFM and Machine Learning”, at the *Bruker and Tufts University BioAFM Workshop*, April 10 2024 in Medford, MA, USA.
- **Talk:** “Identification of geometrical features of cell surface responsible for cancer aggressiveness: Machine learning analysis of atomic force microscopy images”, at the *2024 MRS Fall Meeting & Exhibit*, March 5-10 2024 in Las Vegas, NV, USA.