Towards Concept-based Interpretability of Pre-miRNA Detection using Convolutional Neural Networks

van den Brandt, Irma

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science
Data Mining Research Group

Towards Concept-based Interpretability of Pre-miRNA Detection using Convolutional Neural Networks

Master's Thesis

Irma van den Brandt

Supervisors

dr. Vlado Menkovski (TU/e)
Hilde Weerts, MSc (TU/e)

Assessment Committee

dr. Vlado Menkovski (TU/e)
dr. Natalia Sidorova (TU/e)
Prof. dr. Jens Allmer (Hochschule Ruhr West)
Hilde Weerts, MSc (TU/e)

Eindhoven, October 2021

Abstract

Precursor microRNA (pre-miRNA) sequences are the precursors of microRNAs (miRNAs), which are non-coding RNA sequences regulating gene expression in organisms. Deregulated miRNAs in humans are linked to various diseases, such as cancer and Alzheimer's disease. Given their important role in organisms, domain experts focus on discovering new sequences and researching their functioning. Interestingly, precursor microRNAs can aid the discovery of microRNAs. Also, computational methods have proven to be very successful for detecting (pre-)miRNAs. Contrary to traditional computational methods requiring hand-engineered features to solve the task, Cordero et al. [16] used a model that automatically derives features from the sequences after having encoded them as images. Although this model can successfully perform the pre-miRNA detection task, its practical utility is low due to the limited understanding of the provided predictions. Especially since domain experts desire to use information derived from the predictions to advance their knowledge on structural (and functional) (pre-)miRNA characteristics.

We address this issue by analyzing the usefulness of concept-based interpretability of the pre-miRNA detection results for domain experts. More specifically, by defining concepts based on existing domain knowledge of structural (pre-)miRNA characteristics, we assume concept-based explanations for the predictions can support understanding them. We generate pre-miRNA class predictions using the intrinsic concept-based interpretability method *Concept Whitening* [13], which simultaneously learns solving a classification task and concepts, such that the concepts support explaining the predictions. Moreover, we apply adaptations to the method to support direct quantification of the influence of concepts on the pre-miRNA class predictions.

The main contribution of this thesis is the first step towards concept-based interpretability of the pre-miRNA detection predictions. We propose a framework that aims to provide domain experts with an understanding of a CNN's class predictions for the image-based pre-miRNA detection task in terms of concepts defined based on existing structural pre-miRNA knowledge. In this way, the framework allows confirmation of this existing pre-miRNA detection knowledge. Moreover, domain experts may derive new knowledge from the residual information in the encoded sequences the model considers relevant for the detection task. Although a domain expert raised relevant concerns on the suboptimal understandability of the provided concept-based explanations limiting the practical utility, the framework can be considered a basis for a tool supporting the pre-miRNA detection research.

Contents

Chapter 1

Introduction

For several years, machine learning (ML) systems have proven their success in many domains, such as healthcare and finance. Inherent to their increasing popularity are concerns about their reliability and trustworthiness, and consequently, their practical utility. Especially given their growing use in high-stakes decision applications and their seemingly ever-increasing complexity. Providing system users an understanding (or interpretation) of the system's decisions may alleviate these concerns. With this understanding, different objectives related to the concerns may be realized. For instance, explanations on the system's successful solving of complicated (scientific) problems like analyzing genetic structures in organisms provide domain experts with an intuition of the system's learning, potentially advancing their knowledge in the respective research field.

1.1 Research Problem

In this work, we address the problem of interpreting a given ML model's predictions on the precursor microRNA (pre-miRNA) detection task.

(Precursor) MicroRNAs Pre-miRNAs are hairpin-shaped, non-coding RNA sequences approximately 80-200 nucleotides (nt) in length. *Precursor* miRNAs transform into *mature* microRNAs (miRNAs), which act as gene expression regulators in many biological processes of organisms. Also, deregulated miRNAs are linked to several diseases, such as cancer and various autoimmune, neurological, and cardiovascular diseases [3, 20]. Furthermore, they can serve as biomarkers for disease detection or as drugs in disease treatments [17].

(Precursor) MicroRNA Detection Due to their high importance in organisms, miRNAs have been extensively studied since their discovery in 1993 [8]. However, to date, many details about their functioning remain unknown. Also, it is still unclear which structural characteristics are required for a non-coding RNA sequence to be considered a miRNA [18]. By distinguishing miRNAs from other non-coding RNA sequences using ML methods, researchers can gain insights into these characteristics. These methods, which require data features derived from the characteristics, can be more successful if pre-miRNAs instead of miRNAs are used. Since pre-miRNAs transform into miRNAs, understanding pre-miRNAs can advance our knowledge of miRNAs. Furthermore, pre-miRNAs are longer than miRNAs (80-200 nt versus 18-25 nt) and contain more structural characteristics, both increasing the likelihood of deriving features potentially useful for the detection task.

Because of the lack of knowledge on discriminative miRNA characteristics, most ML models typically require tens to hundreds of features to obtain accurate performance on the detection task [66]. Since the features are human-crafted, they are error-prone [16]. Moreover, using many features may complicate human interpretation of the their importance for the detection task, leading to a lack of understanding of the model's predictions. Recently, Cordero et al. [16] created a novel

approach that automatically derives pre-miRNA features relevant for their detection. Figure 1.1 illustrates the approach's workflow [16]. First, the non-coding RNA sequences are encoded into RGB images, enriched with information on their structure. Next, a convolutional neural network (CNN) performs the pre-miRNA detection task. The automatic feature detection of CNN's makes them successful in case discriminative features for the learning task are (more or less) unknown.

Interpretability of Pre-miRNA Detection Results The approach, named DeepMir [16], is very successful, as pre-miRNAs are detected from the images with a classification accuracy of $\sim 95\%$, which is state-of-the-art performance. In this approach, saliency maps highlighting image parts important for the detection task explain the predictions. Since biologists require information on the importance of *structural (pre-)miRNA characteristics* to perform further research, this type of interpretation is insufficient. Consequently, without improving the interpretability, the approach's practical utility is low. We address this issue by designing a framework enabling the interpretation of a given ML model's predictions based on discriminative pre-miRNA characteristics.

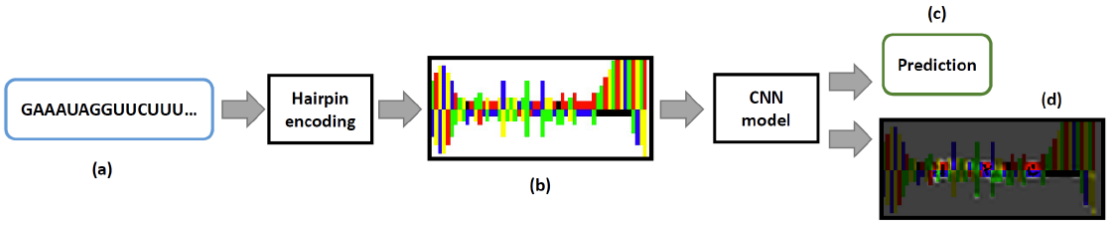

Figure 1.1: Pre-miRNA detection framework of DeepMir [16]. After encoding the RNA sequence as hairpin (a), an algorithm converts it to an RGB image (b). Next, a convolutional neural network (CNN) is trained to perform the pre-miRNA detection task using the images (c). Finally, the CNN's predictions are explained using saliency maps (d).

1.2 Problem Formalization

Based on the context of the research problem above, we can identify the characteristics encompassing the problem. Next, we use them to formulate research questions focusing on the different problem aspects, which together form the research framework of this thesis.

1.2.1 Problem Characteristics

Image-based Pre-miRNA Detection Task First, we define the characteristics concerning the pre-miRNA detection task addressed by Cordero et al. [16].

- *Task*: Binary classification of non-coding RNA sequences into *positive* (or true) and *negative* (or false) pre-miRNAs. We refer to the negative examples as pseudo pre-miRNAs since they can be artificially generated by randomizing positive sequences. As a result, they are not proven to be negative, which complicates the task even further.

- *Data type*. The sequences are encoded as RGB images based on their primary and secondary structure [16]. Figure 1.2 depicts a graphical representation of the two structures of an artificially precursor microRNA designed by Allmer [2]. The primary structure defines the sequence consisting of nucleotides A's, C's, G's, and U's. The secondary structure is derived from the primary one and describes the structural sequence characteristics [6].

- *Goal*: Advancing the scientific knowledge on structural and functional characteristics of (pre-)miRNAs [20]. More specifically, what sequence information is useful for distinguishing between the positive and negative pre-miRNAs.

- *Approach.* The task is solved using ML methods (i.e., *ab initio* pre-miRNA detection, see section 2.2.3).

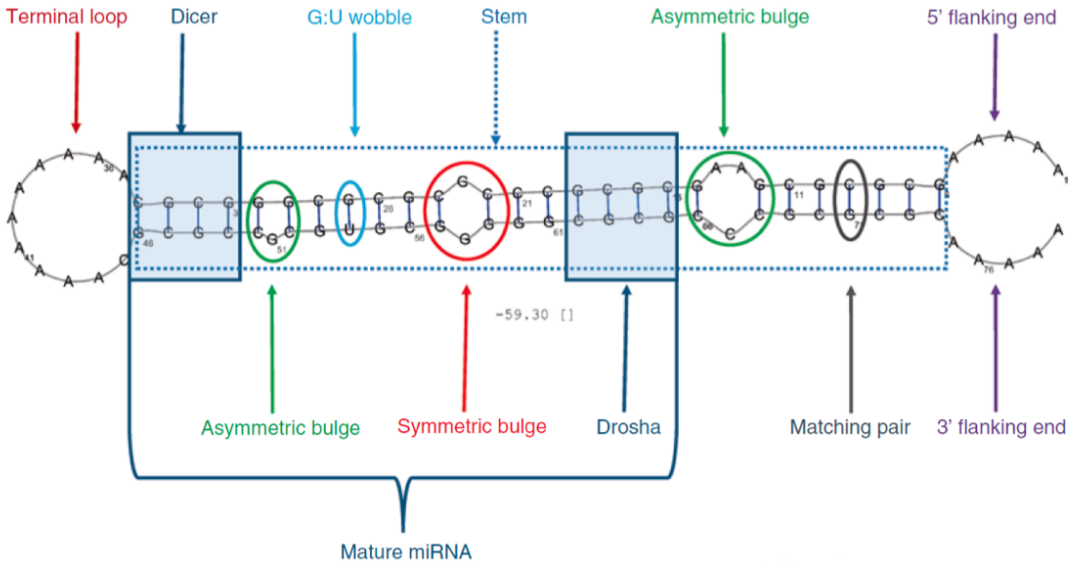

Figure 1.2: Schematic overview of the primary and secondary structure of an artificial precursor microRNA designed by Allmer [2]. The primary structure defines the sequence of nucleotides, the secondary one the structural characteristics such as bugles and loops.

Required Interpretability In general, type, scope, and applicability are factors (see section 2.4.2) that can be used to define the required type of interpretability.

- *Type.* Interpretability methods can be intrinsic or post-hoc. Intrinsic methods generate a self-explainable model, while post-hoc ones explain an existing, trained ML model. Since we are not restricted to use the existing CNN of Cordero et al. [16], both types are possible.
- *Scope.* There are two scopes of interpretability, a local and global one. Local methods inform domain experts on discriminative structural characteristics for a single pre-miRNA, while global ones do this for the input dataset as a whole. In our case, both are important.
- *Applicability.* Interpretability methods can be either applicable for a specific model class (i.e., *model-specific*) or for any model class (i.e., *model-agnostic*). Our interpretability method can be either one of them since there are no constraints concerning the applicability.

1.2.2 Research Questions

Based on the characteristic of the research problem and required type of interpretability, we formulate our main question as follows: *How can we provide domain experts with global and local concept-based explanations of (potentially) discriminative characteristics present in the positive and negative pre-miRNA sequences encoded as images?*

We answer this question using the following three subquestions.

1. *What is the main motivation for requiring interpretability of our image-based pre-miRNA detection application?*

   Several motivations exist for requiring interpretability of an ML system. Also, some interpretability methods are more suitable for particular motivations. Therefore, before choosing a method, we identify our motivation in the context of pre-miRNA detection based on a *literature review* of motivations defined by others requiring their ML system to be interpretable.

2. *Given the context of the image-based pre-miRNA detection problem and the interpretability required to promote discovery of pre-miRNA-related knowledge, which interpretability method is most appropriate?*

   With the *promotion of knowledge discovery* as interpretability motivation, we analyze which method is most suited for the image-based pre-miRNA detection task based on a *literature review* of state-of-the-art ones. We consider methods providing the type of interpretability described in 1.2.1, meaning those with a local and global scope and indifferent type and applicability. Also, it should support identifying discriminative pre-miRNAs characteristics.

3. *How useful are local and global concept-based explanations provided by an interpretability method for supporting domain experts in finding discriminative pre-miRNA characteristics?*

   We use the definition of ML interpretability (section 2.4) to define the usefulness of the interpretability method. This definition states that interpretability means *providing meanings or explanations in a human-understandable manner*. Hence, we consider the method useful if it provides a domain expert with an understanding of its predictions which can support finding discriminative pre-miRNA characteristics. We use the *feedback of a domain expert* on the explanations provided by the method to answer this research question.

1.3 Concept-based Interpretability Approach

From the problem context and formalization, we understand ML-based pre-microRNA detection applications search for features that distinguish *true* pre-miRNAs from *negative* ones. Therefore, we argue the interpretation of these methods is preferably feature-based, as this supports our interpretability motivation to promote the discovery of structural (pre-mi)RNA characteristics. Features can be represented by *concepts*, which are human-interpretable units with a semantic meaning [54, 38]. Consequently, we claim a concept-based interpretability method can support discovering discriminatory pre-miRNA features *and* creates understandable explanations.

Approach Our chosen interpretability method (i.e., *Concept Whitening* [13]), makes an existing convolutional neural network (CNN) interpretable using pre-defined concepts. Namely, the method forces a CNN to learn a classification task and pre-defined concepts simultaneously, such that concept-based (local and global) explanations for the model's predictions can be generated.

To apply this method, we designed a framework consisting of four steps given in Figure 1.3. First, we define concepts based on previous pre-miRNA detection results to increase the likelihood the concepts and concept-based explanations are understandable for domain experts. We refine the definitions based on the presence of and correlation between concepts in the image-based pre-miRNA dataset of interest to increase the concepts' potential to represent discriminative pre-miRNA characteristics. The second step involves training the concept-based interpretability model using the concepts and classification dataset. Finally, we generate concept-based explanations for the model's predictions and evaluate them with a domain expert, whose feedback serves as input for improving the concepts and results.

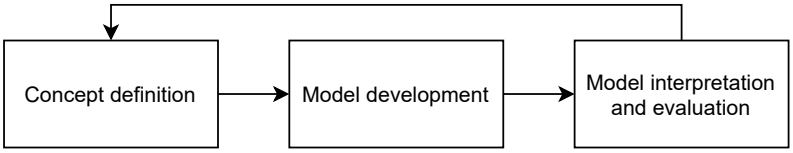

Figure 1.3: The three steps of our interpretable pre-miRNA detection framework. In the *concept definition* step, we create and refine concepts. In the *model development* step, we train our concept-based interpretability model. Finally, we generate concept-based explanations and evaluate them with a domain expert in the *model interpretation and evaluation* step.

Results and Conclusions From the concept-based interpretability framework, we derive the following findings. Firstly, the model's explanations seem to confirm some previous pre-miRNA detection findings. Namely, it finds base pairs in the sequence stem important for detecting true pre-miRNAs from the images. Also, this concept has a negative relationship with the negative pre-miRNA class, further increasing the evidence that it can be discriminative for true pre-miRNAs. Besides this, the model seems to illustrate the importance of an A-U pairs sequence in the (18-25 nt) stem region of pre-miRNAs for positive pre-miRNAs.

Second, the domain expert evaluation demonstrated the usefulness of some framework features in interpreting the pre-miRNA detection results. Our model's ability to combine information learned from pre-defined concepts and residual image information to explain the predictions seems valuable. From the latter, new concepts may be derived. Hence, the framework enables confirming existing pre-miRNA findings and deriving new ones that may advance (pre-)miRNA knowledge. However, the domain expert stressed the current suboptimal understandability of the concept-based explanations limits the usefulness.

Besides the suboptimal understandability, there are some general limitations to the framework. Most important seems the dependence on pre-defined concepts. It is challenging to define useful ones given the limited knowledge on discriminative pre-miRNA characteristics. Also, analyzing their relevance for the pre-miRNA detection task *before* training and evaluating the interpretability model solving the task is complicated. In general, it is unlikely a human creates a concept set encompassing all information required to explain the ML model's predictions. Human concept annotators will probably introduce human bias into the process, leading to a situation where information important for interpretation purposes is overlooked.

1.4 Contributions

The main contribution of this work is a framework for interpretable precursor microRNA (pre-miRNA) detection. As opposed to previous work, our model's explanations rely on domain-inspired *concepts* rather than raw input data. To the best of our knowledge, we are the first to apply concept-based interpretability in the context of the pre-miRNA detection task.

Besides the developed framework, we have made the following contributions. First, we provide a detailed comparison of several state-of-the-art concept-based interpretability methods. Second, building upon the pre-miRNA encoding algorithm of Cordero et al. [16], we define several concepts, leveraging existing (pre-)miRNA detection knowledge. We believe this can provide a basis for future work in this direction. Thirdly, we propose an adjustment to the concept whitening layer allowing direct quantification of the importance of concepts, ensuring faithfulness of the generated explanations. Finally, our model confirms several previous pre-miRNA detection findings, indicating that deep learning models can learn features (more or less) in line with our expectations.

On a final note, we believe future work should focus on defining new concepts and refining the existing ones to improve the interpretable model's performance and explanations. This will bring us one step closer towards intrinsically interpretable, concept-based pre-miRNA detection.

1.5 Outline

The remainder of this report covers the following topics. In Chapter 2, preliminaries related to machine learning, pre-miRNA sequences, the pre-miRNA encoding algorithm of Cordero et al. [16], and ML interpretability are given. Chapter 3 reviews existing literature on interpretability motivations and goals, as well as methods suitable for our required interpretability type. Next, Chapter 4 introduces Concept Whitening (CW) [13], the chosen concept-based interpretability method. Following this method's requirements and concerns, Chapter 5 introduces our framework components. In Chapter 6, we present results obtained using CW for the pre-miRNA detection task and evaluate them with a domain expert. Next, we discuss the framework components and results in Chapter 7. Finally, Chapter 8 covers this work's conclusions and limitations.

Chapter 2

Preliminaries

In this chapter, we introduce several thesis aspects assumed known in the remaining chapters. In section 2.1, the concept of machine learning (ML), and in more detail deep learning, is explained. Next, we provide some presumed information on the precursor microRNA (pre-miRNA) detection task in section 2.2. In section 2.3, we introduce the pre-miRNA image encoding algorithm designed by Cordero et al. [16]. Finally, we introduce the notion of interpretability for ML systems and provide an associated taxonomy used throughout the remainder of the thesis in section 2.4.

2.1 Machine Learning

Machine learning (ML) methods are computational methods that identify data patterns without human intervention. More formally, we consider the prediction problem a given ML method aims to solve by creating an ML model f that learns a mapping from inputs $x_i \in X^m$, with $i \in \{1, ..., n\}$ to outputs Y. Here, m represents the number of *features* and i the number of *instances*. A supervised learning task for model f is $f : X^m \rightarrow Y$ by optimizing parameters θ. The result is a trained model $\hat{f}$ that has learned to predict labels Y of training set $X_{train} \in X$. Hence, for instance $x_i \in X_{train}$ model $\hat{f}$ is trained to predict label $\hat{y}_i$. The training is optimized using a loss function, where the goal is to minimize the loss. The performance of model $\hat{f}$ is typically evaluated by measuring the overlap between predictions made using test set instances $X_{test} \in X$ and their ground-truth labels $y \in Y_{test}$ using an evaluation metric. In unsupervised learning, labels Y are not present. Consequently, unsupervised models f learn the underlying structure or distribution of input data X. Their performance is usually evaluated by analyzing the model's ability to model X's distribution.

ML models can solve difficult problems very well. As the complexity of f increases, more complex patterns can be described, which can result in better predictive performance. However, this leads to the problem that humans cannot understand anymore how model f generates decisions $\hat{y}$ for input instances $x \in X$. These models are often referred to as *black box models*.

2.1.1 Deep Learning

Traditional ML methods require input data X to consist of features useful for the learning task. However, these features are not always as informative. As a result, there exist ML methods that automatically extract useful features from X, such as deep learning (DL) methods. Most DL models consist of a sequence of connected *layers*, such that each one learns some representation of X. This representation is referred to as the layer's latent space [64]. The different layers share their learned information to solve the learning task.

Convolutional Neural Networks Convolutional neural networks (CNNs) are particularly suitable for extracting features or patterns from spatially distributed data such as images. Spatial data is structured such that correlations between features and their relative position to others are

useful for the learning task. CNNs can identify these correlations in a small image region. This, combined with the ability of DL models to learn a hierarchical representation of the input data X through the different layers, makes CNNs very efficient for learning tasks involving spatial data.

CNNs usually contain multiple *convolutional layers* following each other, referred to as hidden layers, followed by one or multiple *fully connected layers*. The convolutional layers identify local patterns in X. The model forwards findings from one layer to the next, which uses them to find new ones. As a result, the model learns increasingly complex patterns, combining them into high-level features such as objects towards the end of the model. Finally, the identified features are mapped to targets Y using the fully connected layer(s). Both layer types include an activation function that maps the output values to a range of values to improve the learning of complex patterns. Therefore, we typically refer to the output of these layers as activations.

Due to the sequential pattern learning, the output of the convolutional layers may increase in size proportional to the layer's location. This can place a burden on the computational efficiency of the model. One way to solve this is to use pooling layers, which sample (or *pool*) values from the layer's activations. There are several ways to do this, such as *Mean* or *Max Pooling*, where we sample the mean or maximum activations, respectively.

Similar to other ML methods, CNNs can have difficulties with generalizing to unseen data, meaning they *overfit* on training data X_{train}. However, one can add several layers that stabilize the training and increase the model's generalizability. For instance, *batch normalization layers* standardize the batch of input data to minimize differences in batches [33]. Alternatively, *dropout layers* that randomly convert a fraction of the layer's activation values to 0 during training could be used [23]. This prevents nodes of consecutive layers to become highly dependent on each other.

Generative Models Generative models are DL models that learn a distribution of input data X. They assume a set of ground-truth factors of variation that form the data distribution can describe input data X [7, 53].

The two of the most common types of generative models are variational auto-encoder (VAEs) and generative adversarial networks (GANs). VAEs consist of an encoder and a decoder, where the first learns a latent representation Z of input X and the latter learns to reconstruct X from Z [43]. These models can be used to manipulate X, and the resulting reconstructions can help analyze the effect of these manipulations. GANs learn a distribution of X using a generative and a discriminative model, G and D, respectively. G generates candidates for the representation of X, while D evaluates them. Additionally, we train G to fool D to improve the quality of D.

2.2 MiRNomics

The precursor microRNA detection task considered in this thesis is part of a research field in bioinformatics concerned with the structural and functional characterization of microRNA sequences. This field is referred to as *mirnomics* by Erson and Petty [20]. MicroRNAs (miRNAs) are non-coding, single-stranded RNA sequences that consist of ∼18-25 nucleotides. Ribonucleic acids (RNA) are a type of nucleic acids that are vital for several biological processes. Before extensively explaining the (precursor) microRNA detection task, we introduce the biogenesis of (precursor) microRNA sequences.

2.2.1 MicroRNA Biogenesis

The biogenesis pathway of most microRNAs consists of the following main steps. First, a long ribonucleotide sequence is created from a transcriptional process involving the miRNA gene. This sequence can consist of several hundred nucleotides and is called the *primary microRNA* (pri-miRNA). In Figure 1.2, a graphical representation of the primary and secondary structure of an artificially designed pre-miRNA created by Allmer [2] is given. The non-coding RNA sequence consists of the 3' and the 5' strands located at the bottom and upper half, respectively.

In the next step, the enzyme Drosha processes the pri-miRNA to generate the *precursor microRNA* (pre-miRNA), a hairpin-shaped, ∼80-200 nucleotides long RNA sequence. Figure 1.2 shows that Drosha cleaves the pri-miRNA such that the left part remains, which is the pre-miRNA. The figure also illustrates the hairpin-like shape of the pre-miRNA. Afterwards, the pre-miRNAs are transported to the cytoplasm, where they are processed by the enzyme Dicer. This process cuts away the terminal loop of the pre-miRNA. Figure 1.2 illustrates the location where the enzyme of interest, Dicer, binds to the pre-miRNA. The result is a miRNA duplex of ∼18-24 nucleotides.

Next, one of the RNA strands is chosen to become the *mature miRNA* and is assembled into the RNA-induced silencing complex (RISC). The other strand of the RNA-duplex is dissolved. Finally, the mature miRNA and its target mRNA start interacting via protein Argonaute and induce the gene expression regulation by reducing the amount of protein synthesized from the target mRNA [29, 49].

2.2.2 Primary and Secondary Structure of (Precursor) MicroRNAs

The primary structure of non-coding RNAs defines the nucleotides sequence. It can contain four different nucleotides, namely adenine (A), cytosine (C), guanine (G), and uracil (U) [6]. The secondary structure contains information on the sequence structure. In Figure 1.2 a graphical representation of the primary and secondary structures of an artificial pri-miRNA and the succeeding pre-miRNA and mi-RNA are illustrated [2]. The representation includes all possible structural characteristics. The two sequences strands, the 3' and the 5' strand, form a nucleotide sequence. They connect via the bonding of their nucleotides. The bond strength is defined by the thermodynamic stability between the two nucleotides [6]. C-G bonds are the strongest, followed by A-U bonds. Both bond types are referred to as *base pairs* (or Watson-Crick pairs) [2]. *Wobbles*, which are G-U bonds, are weaker than A-U bonds. The remaining bond types have a similar strength. Also, two nucleotides of the same type form a mismatch. A mismatch also occurs in the case of a missing nucleotide on a bonding location. Linking this information to the structural characteristics in Figure 1.2, we can conclude that mismatches result in (a)symmetric bulges.

2.2.3 (Precursor) MicroRNA Detection

There exist several other non-coding RNA sequences in organisms besides miRNAs [2]. The goal of the microRNA detection task is to distinguish miRNAs from the other sequences. The task is often executed using *precursor* miRNAs due to the assumption that their longer length in terms of the number of nucleotides and the presence of more structural characteristics simplifies the task [18]. There are two main approaches for the task, *homology-based* and *ab initio* [2].

Homology-based methods detect (pre-)miRNAs based on the similarity between the primary and secondary structure of a new sequence and a validated (pre-)miRNA. The approach assumes homologs of the latter may exist in related species. Notably, the reliance on similarity implies that completely novel (pre-)miRNAs from unrelated species will not be discovered. Since the *ab initio* (pre-)miRNA detection approach considers features commonly derived from the primary and secondary structure characteristics and the thermodynamic information to perform the detection task, it can potentially find completely new (pre-)miRNAs. This approach is also referred to as the ML approach, as it performs the task using computational methods.

2.3 Pre-miRNA Image Encoding Algorithm

In the original DeepMir application, true pre-miRNA sequences are distinguished from other non-coding ones with a CNN trained using image-based examples of both classes. These RGB images are generated using an encoding algorithm created by Cordero et al. [16]. It uses the primary and secondary structure and the thermodynamic information of a sequence to produce the images. Since interactions or bonds between nucleotides and the associated bond strength are considered important for pre-miRNAs, the algorithm incorporates these elements into the image.

The three-step encoding procedure is illustrated with a working example in Figure 2.1 created by Cordero et al. [16]. First, the secondary structure of the RNA sequence is predicted from the primary one using *RNAFold* [32]. The top row in Figure 2.1 (a) represents the primary sequence; the bottom row the secondary one. Brackets denote a matched nucleotides pair, dots a mismatched one. The red section defines the terminal loop of the sequence (reference Figure 1.2). Second, the two sequences are aligned according to the brackets in the secondary one (Figure 2.1(b)). Finally, the aligned sequence is transformed to an RGB image, as shown in Figure 2.1(c). Colored bars represent nucleotides; nucleotide G is red, C blue, U green, and A yellow. The black-colored bars denote gaps. The bar lengths represents the bond strength between two nucleotides. The smaller the length, the stronger the bond. Additional strength is determined by neighboring bonds; if a neighboring bond is a mismatch or a gap, the given one is weakened by adding a pixel to the bars.

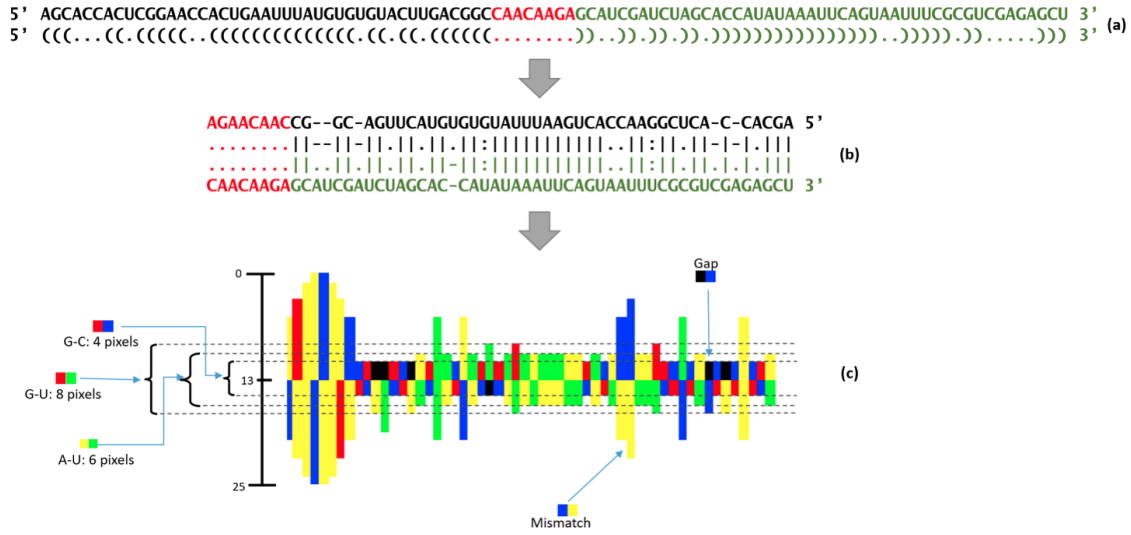

Figure 2.1: Working example of the pre-miRNA encoding procedure of Cordero et al. [16]. In (a) the secondary structure is predicted using *RNAFold* [32]. The top row represents the primary sequence, the bottom the secondary. In (b) they are aligned using the brackets in the secondary one. In (c) the aligned sequence is translated to an RGB image.

2.4 Machine Learning Interpretability

2.4.1 Definition

At this time, the research community has not yet reached a consensus on the definition of interpretability in ML [51]. Nevertheless, the definitions proposed so far agree that interpretability, in general, refers to the action of explaining or providing meaning in understandable concepts [28, 19]. For interpretability in ML, the focus doing this in a *human-understandable* manner [28, 1, 37, 63]. These explanations can be viewed as an interface between the ML system and a human observer [28]. Since interpretability is a domain-specific concept, the community faces challenges in creating a definition that suits all possible purposes of ML applications [63]. Hence, improved definitions should be created with a particular domain and associated knowledge in mind.

Interpretability versus Explainability The terms interpretability and explainability are often used interchangeably in the ML literature. However, a clear distinction between the terms can be made [5, 63, 57, 9]. For instance, Arrieta et al. [5] argues interpretability is a *model trait* expressing the human understandability of a model. Explainability is a *model procedure* used to clarify an uninterpretable model. Hence, explainability aims to provide interpretability [14, 57].

2.4.2 Factors of Interpretability

There are three important factors often used to create a taxonomy on interpretability methods for ML [59, 11]: the *type*, the *scope*, and the *applicability* of the methods.

Type In general, there are two ways to confer interpretability of ML systems, namely intrinsic and post-hoc [51]. Intrinsic interpretability is provided by models that are interpretable by design (i.e., a *white box models*). They are constrained such that they explain themselves in a human-understandable manner [11, 28]. This constraining can degrade the model's predictive performance Lipton [51]. Simpler models are generally more interpretable but may fail to model complex relationships (i.e., the accuracy and interpretability trade-off). However, intrinsically interpretable models have proven to be highly accurate [12, 50, 58, 13, 44]. Also, the added interpretability can support model improvement.

Post-hoc interpretability is given by explaining an existing (black box) model *after training* using some explanation technique. Post-hoc methods approximate a model using an interpretable method explaining the model's predictions [51, 70]. The model's predictive performance is not affected since the explanations are created after training [51]. A negative consequence is that it is uncertain whether the approximation-based explanations truthfully represent the explained model.

To summarize, intrinsic interpretability methods generate interpretable ML models, while post-hoc methods analyze ML models using explanations [38].

Scope The scope of interpretability refers to the amount of information concerning the ML model's predictions process that is explained. It can be local, global, or both. With a local scope, the ML model's decision for one particular instance is explained. Global interpretability methods provide explanations that hold for all input instances or all instances of a single target class. Importantly, aggregating local explanations provides global interpretation [11].

Applicability Interpretability methods can have different application levels. A model-specific level means the method is applicable to one ML model class. In general, methods based on model internals are model-specific. If the applicability is not constrained to one model class, the level is model-agnostic [59]. These methods are not based on model internals since they cannot access the internals by definition. As a result, intrinsically interpretable models cannot be model-agnostic. Model-agnostic methods explain the relation between input and output without considering the model processing input to output [11].

2.5 Summary

In this chapter, we explained that machine learning methods solve a learning task by deriving data patterns mapping input to output. Deep learning models can derive more complex patterns by learning non-linear mappings. This makes them very powerful but complicated to understand.

Afterwards, we introduced precursor microRNAs (pre-miRNAs), which are the *precursors* of microRNAs (miRNAs). MiRNAs play a vital role in several biological processes of organisms. With the (pre-)miRNA detection task, researchers aim to recover information on the functional and structural characteristics of the sequences. ML methods are suitable for solving the task.

Next, we described the pre-miRNA encoding algorithm of Cordero et al. [16].

Lastly, we defined the concept of interpretability of machine learning systems as providing meanings or explanations of the system in a *human-understandable* manner. Furthermore, we introduced the *type*, *scope* and *applicability* factors useful for creating a taxonomy of interpretability methods. The type can be either *intrinsic* or *post-hoc*. Intrinsic methods are interpretable by design, whereas post-hoc ones explain decisions of another model. The methods can have a *global* or a *local* scope. Global methods provide interpretability of the ML system as a whole, while local methods focus on one data instance only. Finally, a method is *model-specific* if it is applicable to one class of ML models, while *model-agnostic* methods are applicable to all model classes.

Chapter 3

Literature Review

In this chapter, we discuss three aspects of interpretability for machine learning systems. In section 3.1, we review different *motivations* for interpretability and analyze which is most suited for our pre-miRNA detection application. Next, we discuss the *goals* for interpretability methods in section 3.2. Since their importance depends on the used interpretability method, we can define which ones are important after discussing several *methods* of interest for our application in section 3.3.

3.1 Motivations for Interpretability

As explained in section 2.4, interpretability of machine learning (ML) systems means to provide a user with meanings or explanations of the system in a human-understandable manner. Interpretability is required in situations where providing users with merely a metric that specifies the model performance is not sufficient [11, 19, 59]. Generally speaking, these are situations that deal with an incomplete problem definition or formulation. This incompleteness leads to missing information that cannot be encoded into the ML model [9]. By interpreting the model, the understanding of the incomplete problem may advance, and issues stemming from the incompleteness may be (partially) solved [11]. Notably, not machine learning (ML) systems require interpretability. There exist situations where high performance is the only demand for the system. In those situations, faulty decisions made by a system have no significant impact. Or the problem the system solves is studied and validated so extensively that the system is trusted completely.

Interpretability can be a means for coping with difficulties that come with solving an incomplete problem [9]. There exist multiple reasons for these difficulties that come with formalizing and defining a problem. The following reasons are the most common ones [5, 11, 19].

- *To provide trust and social acceptance.* A lack of understanding of the decision logic of an ML system can decrease the trust in the system. However, explanations of the decision outcome and associated confidence can increase user trust and acceptance [5, 67, 63, 10].
- *To promote knowledge discovery.* In case one aims to obtain novel (scientific) insights from data, the interpretability of an ML system or its decision function is needed [44]. For instance, the scientific value of a system's output can be validated through interpretability [62]. Also, the interpretations can enable the generation of new hypotheses on the problem at hand.
- *To assess the system's fairness or presence of bias.* An ML system should not discriminate between protected societal groups. However, encoding a fairness definition into the system may be too complicated. Also, the input dataset can include biases against these protected groups [19]. As a remedy, interpretability may help to judge the system's fairness [5].
- *To assess the system's reliability.* Many ML models are unable to generalize across different situations of a problem. Creating a list with all potential failure states of a system is generally infeasible, leaving this behavior undetected [19]. However, we want our system to be reliable

and not generate substantially different outcomes given relatively small input data changes. Interpretability may help in detecting high impact changes [11].

- *To provide justification.* Every day, more companies, governments, and other parties collect sensitive data that could breach an individual's privacy. Since the enforcement of data protection regulations such as the General Data Protection Regulation (GDPR) [15], data processors are obliged to take into account the interpretability of their systems [5, 19, 11]. Besides, they are required to provide data subjects with an explanation of their system's decision function given the subject's data [26].

- To enable interactivity. Applications promoting interaction between user and ML system can also require interpretability. Through interpretability, the target users can better understand their system and potentially improve it [5].

3.2 Interpretability Goals

There exist different goals for interpretability methods, which depend on the motivation for requiring an ML system to be interpretable and the used interpretability method. The ML community has not yet reached a consensus on these goals Carvalho et al. [11]. However, according to Rüping et al. [65], we can define three different yet connected goals. These are the *fidelity* (or *explanation accuracy*), the *understandability*, and the *computational efficiency* of the interpretability method.

Fidelity Interpretability methods aiming for fidelity should faithfully explain a given model's decision logic. Unfaithful explanations can lead to a decrease in user trust in both the interpretability method and the explained model [72]. Contrary to the method's fidelity in terms of classification performance usually evaluated using an accuracy score, the method's fidelity in terms of providing explanations faithful to the explained model is more complicated to quantify [59].

Intrinsically interpretable models adhere to the fidelity goal by design since they are self-explanatory. This property ensures that the provided explanations are truthful to the model's computation [69]. Post-hoc methods, on the other hand, *approximate* the more complex, opaque model with a simpler one. If this model would only generate faithful explanations, the predictive power of both models on the associated learning task would be equal. Hence, we would no longer need the opaque model since we prefer the interpretable one [63, 21].

We can subdivide the fidelity goal into the notions of completeness and soundness. Completeness specifies the extent to which the interpretability method can explain all possible decisions the opaque model can make given a particular input instance. In other words, it defines the representativeness or generalizability of the interpretability method [72, 69]. A more complete explanation is considered more faithful [25, 48]. The soundness of a method defines the fidelity of the method's elements with respect to the opaque model it explains. The lower the interpretability method's complexity compared to that of the opaque model, the less sound it is considered [47].

Understandability Understandability defines a given user's ability to grasp the explanations of an interpretability method. Faithful explanations not reasonable nor appealing to humans may lose their value [11]. The understandability of interpretability methods is hard to measure due to the vagueness surrounding the definition of interpretability. However, some characteristics of understandability can support the evaluation of the goal. More specifically, Miller [57] identified the following characteristics that define human-friendly understandability based on findings from social sciences.

- *Contrastiveness.* This characteristic originates from the human tendency to think in counterfactual cases. Hence, it focuses on identifying which factors can induce an outcome change. An example is an explanation that illustrates that a bank would offer an application a loan if this person would earn €1,000 more per month.

- *Selectivity.* Explanations should focus on the most important or unexpected factors that cause an outcome. Requiring someone to recover them from a long list of possible factors is undesirable. Moreover, selective explanations can prevent any possible disagreement between different humans. Sometimes selectivity may not be desirable. Using multiple explanations may be beneficial for directing the user to more careful and rational thinking [35].

- *Social.* The social interaction between method (or *explainer*) and audience (or *explainee*) defines, given the audience, the most suitable type of communication for the explanations.

- *Consistency with prior beliefs explainee.* Humans tend to ignore information that contradicts their prior beliefs. This characteristic can be a trade-off with fidelity since explanations may not always be truthful and consistent with prior beliefs.

- *Focus on the abnormal.* Besides focusing on prior beliefs, humans may also prefer unexpected causes when given explanations.

- *Generable and probable.* Contradictory to the previous characteristic, it may be desirable to provide explanations that hold for most cases, especially when abnormal causes are not present [11].

- *Causal.* Humans prefer obtaining causal explanations, as they can explain the underlying cause for a decision or phenomenon [14, 48].

Notably, these characteristics can also be hard to measure. For some of them, we can define clear metrics. For instance, we can measure the *generable and probable* characteristic using the ratio of the number of data points for which the explanation applies and the total number of data points. Other characteristics may require additional knowledge of the application domain and audience.

Computational Efficiency The computational efficiency of an interpretability method refers to the computational resources required by the system to generate explanations. Examples of computational resources are computing power or amount of storage.

Trade-offs There exist several trade-offs between these goals. For instance, in some cases, the more accurate the explanation method is, the less understandable it becomes [25]. Only a complete one can be considered perfectly faithful. However, this requires it to be extensive, potentially decreasing its understandability. Ideally, the two goals are balanced. Constraining a method too much to increase the understandability and fit with user preferences can lead to implicit human cognitive biases in the explanations, which may endanger the fidelity of the method [31].

Also, increasing the extensiveness or complexity of a method to increase its fidelity is likely to decrease its computational efficiency. Formulated differently, approximating an opaque ML system with a simpler model lowers the computational cost but also the level of fidelity.

3.3 Concept-based Interpretability Methods

We will focus our search for interpretability methods for the image-based pre-miRNA detection application on concept-based ones. There are several reasons for this. First of all, Mahinpei et al. [54] argues concepts allow for more meaningful user interaction since they have semantic meaning to the user, as opposed to explanations that highlight (groups of) pixels to quantify their importance (i.e., saliency maps). Moreover, humans often use known concepts to explain their decisions, increasing potential understandability [63, 24]. Indeed, concept-based explanations have been successfully applied in the healthcare domain [45]. Additionally, they seem particularly suitable for the field of miRNomics (which includes the (pre)-miRNA detection task), as we can represent the detected structural and functional features as concepts.

Kazhdan et al. [37] has identified two main approaches for these types of methods. We refer to the first class of methods as *methods using pre-defined concepts to provide concept-based interpretability*. These methods use pre-defined concepts to explain an ML system's decisions. The second class focuses on the automatic derivation of concepts from the input data. More specifically, they learn a disentangled space with disjoint data variation factors using a generative model. We refer to this learning as *disentanglement learning*. These factors (or latent factors of the disentangled data space) can be seen as concepts.

There exist several methods for the two approaches. In Figure 3.1, we subdivide the methods of the two approaches included in this project into smaller categories. For the concept-based approaches, this is based on their type (i.e., intrinsic or post-hoc). For the disentanglement learning approaches, we use the level of supervision included (i.e., (semi-)supervised or unsupervised). Next, we introduce the two approaches and their methods and discuss them further. The discussion is characterized using three properties of interpretability methods, namely their *type*, *scope* and *completeness*. The first two properties are part of the factors of interpretability 2.4.2. The remaining factor of interpretability, the *applicability*, is not included since all methods are model-specific. The *completeness* property is already explained in the section on goals of interpretability methods 3.2.

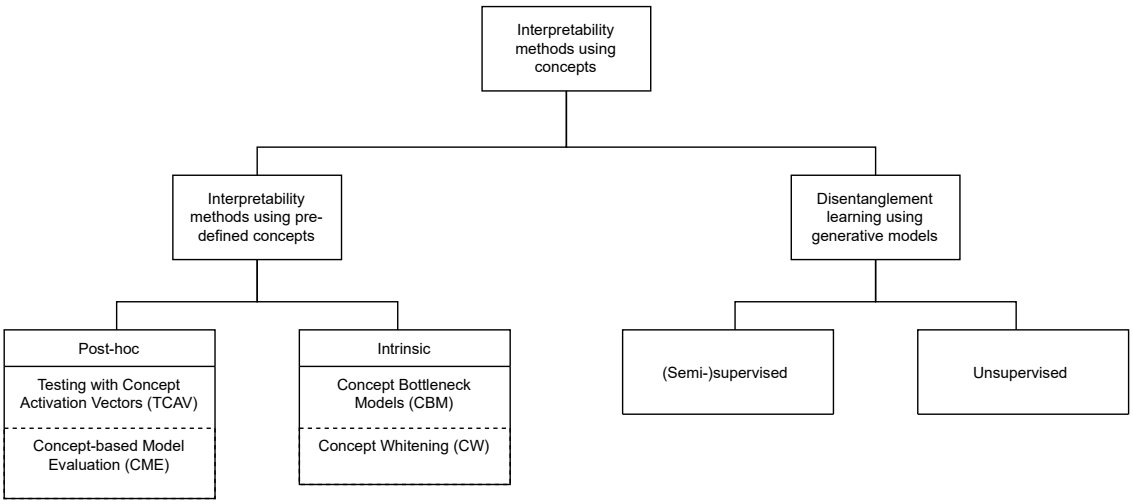

Figure 3.1: Characterization of the interpretability methods using concepts. TCAV [42] and CME [38] are post-hoc concept-based interpretability methods, CBM [45] and CW [13] are intrinsic ones. The learning of a disentangled latent space for interpretability purposes is subdivided based on the level of supervision that is applied.

3.3.1 Interpretability Methods using Pre-defined Concepts

Interpretability methods using (pre-defined) concepts explain a deep neural network's (DNN) decision function using these concepts. They generate explanations by linking the concepts to the internal representations of a DNN [74, 38]. We use an example to illustrate the effectiveness of the methods compared to pixel-based methods. Consider a model trained to classify animals and objects from images. As Rudin [63] explains, pixel-based explanations such as saliency maps can highlight similar image regions irrespective of the class prediction. Thus, the region important for classifying a dog image as a musical instrument can be practically identical to that would the predicted class be a dog. As a result, the explanations do not help in understanding the model's decisions. For concept-based interpretability methods, this is not an issue. They use semantically meaningful explanation units during explanation instead of non-semantic units such as (groups of) pixels.

In Table 3.1, the four concept-based interpretability methods included in this section are characterized using the properties of interpretability methods (i.e., the *type*, *scope* and *completeness*). We see all methods have an equal scope (i.e., both global and local), while their type and completeness can differ. Below, we discuss the application of the three properties in concept-based interpretability methods. Afterwards, we introduce the four methods and compare them.

- **Type**. The type refers to the method being post-hoc or intrinsic. Post-hoc concept-based interpretability methods explain a DNN's decision function with pre-defined concepts by recovering these concepts from the DNN's internal representations. Intrinsic methods interpret the decision function in terms of concepts by adapting parts of the DNN so that these parts induce interpretability. Since this enables the model to learn concepts, they are also referred to as *Concept Learning Models* [54].
- **Scope**. The scope defines the method's applicability domain concerning the size of the input dataset. Global concept-based interpretability methods define the importance of concepts for the entire dataset or a complete target class in the dataset. Local methods focus on a single instance or small subset of the entire dataset when defining the importance of concepts.
- **Completeness**. Completeness defines the ability of the method to truthfully explain all possible scenarios for the DNN's decision function given a set of (pre-defined) concepts. This property is important for methods constrained to use a set of pre-defined concepts since they usually only describe some variation in the input dataset. As a result, the method explains only part of the model's decision logic. If the data generation process is known and simple, it is more likely that it is possible to define a complete set of concepts [38].

Method	Type	Scope	Completeness
TCAV [41]	Post-hoc	Local & Global	No [73]
CME [38]	Post-hoc	Local & Global	No
CBM [45]	Intrinsic	Local & Global	No
CW [13]	Intrinsic	Local & Global	To some extent

Table 3.1: The four concept-based interpretability methods characterized by their type, scope and completeness.

Testing with Concept Activation Vectors The first post-hoc concept-based interpretability approach discussed is Testing with Concept Activation Vectors (TCAV) [41]. The method requires a pre-trained CNN f, a set of concept images P_c, a set of target images X containing the concepts in P_c, and a set of random counterexamples of the concept images N. Concept images P_c and random counterexamples N can be from X or another dataset.

TCAV measures the sensitivity of a DNN's decision function using concept activation vectors (CAVs). These are vectors orthogonal to the decision boundary of a linear classifier trained to distinguish between the internal representations of the concept images $\mathbf{x} \in P_c$ and random counterexamples $\mathbf{x} \in N$. The internal representations $f(\mathbf{x})$ are collected from f at layer of choice l. Hence, CAVs present a learned concept representation at layer l. The derivative of representation $f(\mathbf{x})$ of input image $\mathbf{x} \in X$ in the direction of the CAV defines the sensitivity of layer l to the concept for input instance $\mathbf{x}$. This concept sensitivity score is a local explanation. The fraction of all inputs from one target class positively influenced by the concept provides a global explanation for that class. This procedure is referred to as Testing with CAVs (TCAV).

Instead of requiring pre-defined concept examples (and labels) for TCAV, concepts can also be derived automatically from the input images $\mathbf{x} \in X$. For instance, Ghorbani et al. [24] segments the input images to extract useful concepts from them. Afterwards, TCAV is used to define the importance of the extracted concepts for a pre-trained CNN.

The quality of the CAVs depends on the linear classifier's ability to distinguish between the concept images and random counterexamples. Hence, we require the internal representations of those two sets to be linearly separable from each other. If so, the CAVs can be assumed uncorrelated and representative of a single concept (i.e., they are *pure*) [64, 27]. This is the case when the latent space of the layer including the internal representations are used is disentangled. However, Zhou et al. [75] has shown that assuming standard DNNs to possess disentangled latent spaces is not grounded. Moreover, TCAV cannot impose disentanglement constraints on the latent space of the original DNN due to the post-hoc nature of the method [13, 74, 45]. As a result, we cannot assume disentanglement. Therefore, the ability of the TCAV method to recover concepts in the form of CAVs from a DNN not trained using concepts is questionable. As a result, the concept sensitivity results can be misleading. Finally, since TCAV can only define concept sensitivity for concepts in P_c, the method is not complete if P_c cannot sufficiently explain all variation in the input data X.

Concept-based Model Evaluation (CME) Concept-based Model Evaluation (CME) [38] is a post-hoc interpretability method that combines model extraction techniques with concept-based explanations. The method requires a set of input data X, (partially) labeled with a set of concepts C, and a pre-trained DNN f. First, an *Input-to-Concept*-function predicts concepts $\hat{\mathbf{c}} \in C$ from the internal representations of input data $\mathbf{x} \in X$ at layer l in DNN f. In the case of partial concept labels, some input instances $\mathbf{x} \in X$ are labeled with concepts. From those we can derive ground truth representations $f_l(x)$ of the concepts at layer l. These representations $f_l(x)$ are used to train the *Input-to-Concept*-function with Semi-Supervised Multi-Task Learning (SSMTL), where every concept is considered a different task. Second, a *Concept-to-Output*-function uses concepts $\hat{\mathbf{c}} \in C$ to classify input instances $\mathbf{x} \in X$. This step should be performed with an interpretable classification model, such as a decision tree. As a result, the concepts $\hat{\mathbf{c}} \in C$ predicted from input image $\mathbf{x} \in X$ can be linked to the prediction, giving a local explanation. Global interpretation is provided through testing the concept-decomposability of DNN f using the given concepts C.

The CME approach assumes the concepts in C to be relevant for the classification task. In other words, it considers the opaque DNN to be concept-decomposable with the identified concepts. However, experiments conducted by Kazhdan et al. [38] show this assumption can be too strict. Consequently, the resulting explanations may not be complete. Besides, CME relies on the concept recovery ability of the SSMTL extraction model. Like TCAV, this can lead to misleading conclusions when applied to black-box DNNs that are not constrained to use concepts during training.

Concept Bottleneck Models (CBM) Concept Bottleneck Models (CBM) is an intrinsic interpretability method that, similar to CME, consists of two stages. The method requires a set of input images X, split into a training and test set and labeled with concept set C. First, a DNN f is trained to predict concept labels $\hat{\mathbf{c}} \in C$ from the internal representations of an input image $\mathbf{x} \in X$. This is done by resizing one layer l of f to match exactly with the number of concepts. During training, an intermediate loss optimizes the alignment of the neurons in layer l component-wise with the provided concepts $\mathbf{c} \in C$. In the next step, the model generates class predictions using the predicted concepts $\hat{\mathbf{c}} \in C$. As a result, the predicted concepts can explain the class prediction.

The technique is similar to CME, however, CME applies these steps to *analyze* an existing DNN while CBM uses them to *generate* (or train) a DNN. Consequently, CBM is concept-decomposable by design. Another advantage of CBM compared to CME is that CBM allows for human interventions on the predicted concepts $\hat{\mathbf{c}} \in C$ during testing. This may increase both the classification accuracy of the overall model and the human understanding of the model. Due to this similarity, the local and global interpretability of CBMs is similar to that of CME.

Since CBMs are trained to use concepts only for prediction, concept recovery issues are not present here. However, their explanations can still be misleading. They require the pre-defined concepts to be highly predictive for the classes of the classification task. This does not mean they are causal. They only need to guide the classification model to a preference for a certain class. However, quantifying the predictiveness of a concept for a class is complicated. If concepts are not highly predictive for a class, CBMs can become less accurate than expected, or their explanations

can become misleading. The latter happens when concept irrelevant information present in the concept images is used during prediction [37]. Indeed, Margeloiu et al. [55] have shown that CBMs may not restrict themselves to the learned concept information during class prediction (i.e. *information leakage* [54]). Therefore, we have to conclude that CBMs are not complete.

Concept Whitening The second intrinsic concept-based interpretability method discussed is Concept Whitening (CW) [13]. CW requires an input dataset X, consisting of a training and test set, a set of concept example images C, and a pre-trained CNN f. The set of concept examples can be from X, but this is not required.

The CW approach consists of two main steps. First, the method whitens the latent space of the layer of choice l. This means that the latent space is standardized and decorrelated, which encourages disentanglement. Next, the pre-defined concepts $\mathbf{c} \in C$ are aligned to the nodes in l one by one using an orthogonal transformation. In this way, all information of one particular concept $\mathbf{c} \in C$ flows through one node of l only. As a result, the axes (or nodes) of the latent space of l can be used to interpret the internal representations of l using our concept set C for global interpretability of f. On a local level, one can define the importance of a concept through the activity of the nodes that are aligned with the concept during prediction of input instance $\mathbf{x} \in X$.

The alignment of nodes in layer l and pre-defined concepts $\mathbf{c} \in C$ in an element-wise fashion can lead to a subset of nodes in l not aligned with any pre-defined concepts. These nodes handle residual information left in the input images $\mathbf{x} \in X$ that is not explained by the concept examples $\mathbf{c} \in C$. This can enable the definition of new concepts from the residual information left in X [13]. Since this allows to generalize over decision paths the CNN can deduce from X, CW is considered more complete. However, it may occur these remaining nodes handle information on a concept or something very similar. Consequently, information leakage can still occur [54].

3.3.2 Disentanglement Learning using Deep Generative Models

As concept labels can be difficult to define or obtain, there exist several approaches that automatically derive concepts from image data. These approaches can use a deep generative model that learns the underlying distribution of the data, consisting of statistically independent factors of variation. The learned representation of the input data is referred to as a *disentangled latent space* [64]. The independence between the factors supports interpreting the learned data representation. Namely, the factors represent relationships present between the input data and output data. The learned representation can describe simple data features or more complex patterns [61]. Kazhdan et al. [37] argues that these factors may be seen as concepts in case they are *interpretable by humans* [37]. As such, disentanglement learning might be used for interpretability purposes.

Disentanglement learning can be done with different levels of supervision. In a supervised setting, the user decides which factors are learned [56, 53]. In the case of *complete supervision*, the method can better be characterized as a concept-based interpretability method. Interestingly, the concept-based method *Concept Whitening* [13], discussed in section 3.3.1, is an example of such an approach. However, it does not use a generative model to learn the disentanglement. In an unsupervised learning scenario, on the other hand, the factors are derived automatically from the input data without providing the model with any user-defined knowledge of the factors. As a result, they are not influenced by humans bias, contrary to predefined concepts [24].

In the remainder of this section, we discuss several interpretability methods using deep generative models to learn disentanglement. First, unsupervised methods are discussed, followed by (weakly) supervised ones. Linking them to the methods listed in Table 3.1, we can characterize all disentanglement learning methods as follows. They are intrinsically interpretable, however, they may be combined with post-hoc interpretability methods for a more extensive explanation. The scope depends on the way they generate explanations. All methods can be considered complete since they can automatically derive all factors of variation in the input data.

Unsupervised Disentanglement Learning Since there is no human bias involved in unsupervised disentanglement learning, the learned latent factors may reveal novel and more complex relationships between input and output [69]. They have the potential to advance human knowledge about complex problems, which is the aim of scientific discovery [64].

For instance, O'Shaughnessy et al. [61] uses a generative approach to learn a disentangled space where causal and non-causal factors are separated without any supervision. A factor is causal if a change in the factor induces a change in the output prediction. The causal and non-causal factors together represent the complete data distribution. The number of factors is determined using an algorithm that balances the data fidelity of the learned distribution and the load on the causal influence term. With the factors, causal explanations between images and labels can be generated by changing one or multiple causal factors and keeping the non-causal ones fixed. Global explanations are provided by visualizing the effect of a change in causal factors on the learned data representation. Local explanations are given through analyzing which factors are crucial to obtain a particular class prediction. Importantly, experiments show that interpreting the factors is very difficult. As a result, the explanations can leave the user guessing about what they represent. Similar to the issues with pixel-based interpretability methods, this leads to situations where the explanations do not support the goal of gaining an understanding of the model's decisions.

Interpretation issues concerning the latent factors of a disentangled space learned without supervision are common. Locatello et al. [52] found that unsupervised disentanglement learning can result in finding infinitely many factors that represent the input data very well but are not guaranteed to be understandable. Hence, using an unsupervised approach seems to be inconvenient given the goal of generating human-understandable explanations [37].

(Semi-)supervised Disentanglement Learning Using some supervision while learning a disentangled latent space can improve the interpretability of the space [53]. Namely, the added supervision can support in finding the meaning of the learned factors of variation.

For example, Locatello et al. [53] created a weakly supervised approach where a generative model learns disentanglement based on input image pairs that differ in some factors of variation. Hence, users should have some assumptions about the data variation to create these image pairs [37]. Explanations describing the relation between input and target class can be generated as follows. The approach uses a variational auto-encoder that consists of an encoder and a decoder. With the decoder, one can make reconstructions of the original input images. They can illustrate the effect of changing one or more latent factors on the data representation. Moreover, by providing a reconstruction to a classification model and analyzing whether the class prediction changes with respect to the original one, one can make counterfactual explanations.

Perturbation-based explainability methods commonly used for low-dimensional data can also be applied. They interpret a model's decision function by changing input factors and measuring the influence on the output. Hence, they are computationally expensive for high-dimensional data such as images. However, the computational burden significantly decreases when using the learned latent factors, which represent a compressed version of the high-dimensional data. For example, Mijolla et al. [56] created a framework that uses Shapley values based on the learned latent factors for generating explanations. Shapley values measure the average contribution of a feature value to the output of a model to provide a local explanation [59]. Global explanations can be computed by averaging the local values. Note that different levels of supervision can be applied in the framework, however, unsupervised learning seems undesirable given the possible interpretability issues.

Learning a disentanglement data representation in a supervised manner enables validating whether the assumed pre-defined factors of variation are indeed factors of variation. Besides, other factors handling the remaining variation are learned automatically. However, interpreting the learned factors may be challenging, especially with more complex input datasets. Additionally, the approaches assume that disentanglement can be learned. As Rudin et al. [64] argues, this can be challenging and hard to evaluate, especially with lower levels of supervision.

3.3.3 Summary

In this section, we first defined that interpretability methods using *human-understandable concepts* are desired for interpreting the pre-miRNA data. The structural and functional characteristics of (pre-)miRNAs that are our interest may be seen as concepts. Moreover, human-understandable concepts are considered a meaningful way to explain a DNN.

With the focus on concepts, we introduced two concept-based interpretability approaches. The first one aligns pre-defined concepts to a DNN's internal representations. This can be done post-hoc or intrinsically. Post-hoc methods assume that the DNN they explain has learned the concepts of interest to perform the alignment. However, the methods do not prove this assumption, making it not grounded. As a result, their explanations can be misleading [45, 13, 64]. The intrinsic methods aim to solve this issue by forcing a DNN to learn concepts.

The other approach for providing concept-based interpretability focuses on learning a disentangled latent space using a generative model. This space contains independent factors of variation derived from the input data. If they are human interpretable, they may be seen as concepts. Hence, the main difference with the previous approach is that concept labels are not required since the model learns them *automatically*. However, interpreting these derived concepts can be complicated and sometimes even impossible due to their high level of complexity. Creating uninterpretable concepts is highly undesirable given the understandability goal of interpretability methods.

3.4 Conclusions

In this section, we discussed several aspects of interpretability for ML systems. We concluded that interpretability is relevant when solving a problem with an incomplete definition or formulation, which can stem from various reasons. For the pre-miRNA detection problem, the reason for interpretability is to *discover (scientific) knowledge* from the ML system's outcomes. The interpretability may support advancing the knowledge on (pre-)miRNAs.

Next, we found that there are three main goals of interpretability methods, namely *fidelity*, *understandability* and *computational efficiency*. The importance of these goals for the pre-miRNA detection problem depends on the type of interpretability method chosen.

Finally, we defined the desire for local and global concept-based interpretability for our application. We introduced two approaches providing this type of interpretability, the *interpretability methods using pre-defined concepts* and the *disentanglement learning using deep generative models* approaches. We found the latter undesirable given that the automatically learned concepts can be uninterpretable. The definition of interpretability (section 2.4) emphasizes the requirement of *human-understandable* explanations. From the approaches using pre-defined concepts, we conclude that the inherent *fidelity* of intrinsic concept-based interpretability methods makes them most appropriate to interpret the pre-miRNA detection results. Since *concept bottleneck models* [45] assume all pre-defined concepts to be relevant for the original classification task, which is generally not the case, their explanations can be misleading. As a result, they can be incomplete. Since the *concept whitening* method [13] uses residual information from the input data besides pre-defined concepts during training, it is more complete. Therefore, we choose the *concept whitening* method as our interpretability method. The method's inherent *fidelity* enables us to focus on its *understandability*, which we require to support the search for discriminative structural pre-miRNA features.

Chapter 4

Background Information

In this chapter, we introduce the concept-based interpretability method, named *Concept Whitening* (CW) [13], chosen based on the literature review in Chapter 3. The method makes an existing convolutional neural network (CNN) interpretable using pre-defined concepts [13]. The CNN is adapted to learn concepts and their link with the classification task of interest. Consequently, we can use the concepts to explain the class predictions. In section 4.1, we provide a detailed description of the technique. In section 4.2, we describe which explanations can be generated. We end with some concerns regarding these explanations in section 4.3.

4.1 Method

The concept whitening technique is applied to one or multiple layers of an existing CNN. The method aims to learn a *meaningful* latent space in terms of pre-defined concepts in these layers in two steps. First, it disentangles the space using standardization and decorrelation. Then, it rotates it so that each pre-defined concept is linked to a different axis of the multidimensional space. Consequently, the axes can be used to interpret the model's concept learning and to investigate to what extent input images contain a concept.

Formulating the method formally, consider image dataset X and labels Y. Also, we have a concept image dataset C. Like other image classification algorithms, the method aims to learn a mapping $f : X \rightarrow Y$ to classify images X with labels Y by minimizing a loss function (e.g., (binary) cross-entropy loss) such that the classification accuracy metric is maximized. In the case of CW, f is a CNN. Contrary to standard CNNs, CW aligns each concept $c_j \in C$ with $j \in \{1, .., k\}$ with the j^{th} axis of high-dimensional latent space Z. Hence, the model's representations of concept c_j, represented by a cloud of points in space Z, are structured such that they are located close to or on top of axis j of Z. In case space Z has more axes m than pre-defined concepts k, thus $m > k$, the remaining $m - k$ axes are left unaligned.

Next, we explain the two main steps of the concept alignment procedure, namely the whitening transformation (1) and the orthogonal transformation (2), in more detail.

4.1.1 Whitening Transformation

To align concepts with latent space axes, latent space Z should be disentangled. The CW method forces disentanglement by applying a *whitening* transformation on Z. This linear transformation transforms a vector of random variables with a known positive definite covariance matrix into a new random vector such that the covariance matrix is equal to the identity matrix [39]. In our case, each column $\mathbf{z}_i \in \mathbb{R}^d$ of latent space representation $\mathbf{Z}_{d \times n}$ of layer l consists of latent features for instances $x_i \in X$ for $i \in \{1, ..., n\}$. The whitening transformation applied on $\mathbf{Z}_{d \times n}$ is as follows:

$$\phi(\mathbf{Z_{d \times n}}) = \mathbf{W}(\mathbf{Z_{d \times n}} - \mu \mathbf{1}^T_{n \times 1}) \tag{4.1}$$

with sample mean $\mu = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i$ and whitening matrix $\mathbf{W}_{d\times d}$. After having applied transformation $\phi(\mathbf{Z}_{d\times n})$, every axis of space $\mathbf{Z}_{d\times n}$ is standardized and decorrelated. See appendix A for a more detailed explanation of whitening transformations in general. Figure 4.1 illustrates the effect of a whitening transformation to a latent space. In Figure 4.1(a), the points in space are correlated with each other and not standardized, in (b), they are standardized but still correlated, and in (c), they are standardized and decorrelated (i.e., they are *whitened*).

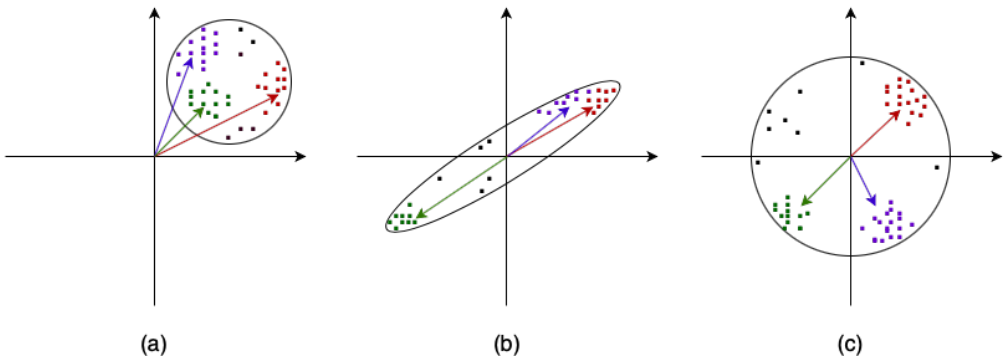

Figure 4.1: Graphical representation of a two-dimensional latent space. In (a), the space is correlated and not standardized. In (b), it is correlated and standardized. In (c), the space is whitened, forcing decorrelation and standardization simultaneously. The arrows denote the directional derivatives towards similar points in the space.

4.1.2 Orthogonal Transformation

Following the whitening transformation of space Z is an orthogonal one, which enables aligning concepts $c_j \in C$ with the axes of $\mathbf{Z}_{d\times n}$. In this way, the axes can be used to analyze the model's concept representations and their importance for the underlying classification task. The transformation rotates the whitened latent space $\phi(\mathbf{Z}_{d\times n})$ such that latent space representations $\mathbf{Z}_{c_j} \in \mathbf{Z}_{d\times n}$ of concept samples $\mathbf{X}_{c_j}$ are maximally activated on latent space axis j. The orthogonal transformation is performed using a d-dimensional orthogonal matrix $\mathbf{Q}$, where column $\mathbf{q}_j$ represents axis j. Matrix $\mathbf{Q}$ is optimized using objective function:

$$\max_{\mathbf{q}_1,\ldots,\mathbf{q}_k} \sum_{j=1}^{k} \frac{1}{n_j}\mathbf{q}_j^T \phi(\mathbf{Z}_{c_j})\mathbf{1}_{n_j\times 1} \quad \text{such that} \quad \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_d \tag{4.2}$$

The orthogonal matrix $\mathbf{Q}$ is learned through optimizing objective 4.2 using gradient descent. Matrix $\mathbf{Q}$ and classification mapping $f : X \to Y$ are optimized in turns. Hence, $\mathbf{Q}$ is fixed when optimizing the classification loss, while the parameters of the classification objective remain fixed during concept alignment. Appendix A contains a general description of orthogonal transformations. In Figure 4.2, an example of a latent space before (a) and after (b) an orthogonal transformation is given. In Figure 4.2(a), the concept representations are separated from each other but not aligned with the axes. In Figure 4.2(b), the representations are rotated such that they (almost) align with the axes. With this space, we can use the axes to interpret the space and its disentanglement.

The rotation of space $\mathbf{Z}_{d\times n}$ using rotation matrix $\mathbf{Q}$ should ensure that latent space representations $\mathbf{Z}_{c_j} \in \mathbf{Z}_{d\times n}$ of concept samples $\mathbf{X}_{c_j}$ are maximally activated at axis j and concept examples not in $\mathbf{X}_{c_j}$ (i.e., examples of different concepts) are not. The optimization algorithm learning $\mathbf{Q}$ expects concept activation scalars instead of concept activation maps (or matrices). However, the information on the activation level of an input image part by filter (or axis) j aligned with concept c_j is stored in map $\mathbf{Z}_{c_j} \in \mathbf{Z}_{d\times n}$. Therefore, CW reshapes every map $\mathbf{Z}_{c_j}$ into a scalar. [13] found shaping using the mean of all max-pooled values generally gives the best concept alignment results since this combines both low- and high-level image information.

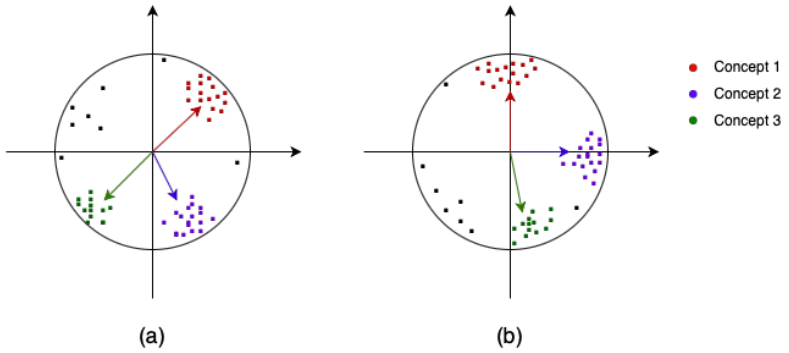

Figure 4.2: Graphical representation of the effect of an orthogonal transformation applied on a (whitened) two-dimensional latent space. The transformation encourages concept representations to be aligned with axes of the latent space.

Down-sampling the maps using max-pooling captures high-level objects located in specific image locations. Low-level features, which are more widespread in the image, are recognized by taking the mean of the values. Depending on the layer location, the procedure focuses more on either high- or low-level features. Towards the end of the network, the max-pooling becomes more dominant.

4.2 Generated Explanations

A CNN applying the CW method is forced to be interpretable, enabling the model to explain itself [5]. Indeed, it can provide information on the learning of pre-defined concepts and their importance for the prediction task. Since these explanations depend on the independence of the learned concept representations, the latent space consisting of these representations should be properly disentangled. The extent to which a learned concept solely represents the original concept is known as the concept *purity*. Importantly, all concept-based interpretability methods strive for pure concepts to measure their influences on the predictions without suffering from correlations with others [64], as this can result in misleading explanations.

Next, CW's way of analyzing the different explanation aspects (i.e., *concept learning*, *concept importance* and *concept purity*), are introduced. For completeness, we also explain the *concept representations* provided to an explainee.

Concept Learning We can obtain an interpretation of the model's internals in terms of concept learning by visualizing the maximally activated images for concepts $c_j \in C$ on latent space axis j. They illustrate how concept c_j is represented by model f. Additionally, the empirical receptive field of axis j, referred to as the receptive field of concept c_j after concept alignment, can be added to the visualization to clarify the focus of the filter that represents this axis (or concept) [4].

Concept Importance The importance of the learned concepts $c_j \in C$ can be defined through investigating whether latent space axis j aligned with concept c_j is (highly) active during prediction of input instance $x_1, ..., x_n \in X$. The activity defines the importance of concept c_j to some extent.

Concept Purity There are two different ways to analyze the pureness of concepts. The first measures the inter- and intra-concept similarity. The inter-concept similarity is calculated based on the average pairwise cosine similarity between latent concept representations of two different concepts. The formal definition is as follows:

$$d_{ij} = \frac{1}{nm}\left(\sum_{k=1}^{n}\sum_{l=1}^{m}\frac{\mathbf{Z}_{c_i k} \cdot \mathbf{Z}_{c_j l}}{||\mathbf{Z}_{c_i k}||_2 ||\mathbf{Z}_{c_j l}||_2}\right) \tag{4.3}$$

with n and m the number of instances for concept c_i and c_j, with $i \neq j$. The representation of concept c_i and instance n is denoted by $\mathbf{Z}_{c_i n}$. The lower the similarity score, the more dissimilar the representations of two different concepts. The intra-concept similarity, which computes the similarity between representations of the same concept, is defined likewise. These scores should be 1 to illustrate the similarity of representations of the same concept.

The second approach computes a one-vs-all AUC score to measure the distinguishability of the representations of different concepts. For all exemplary images $\mathbf{X}_{c_j}$ of all concepts $c_j \in C$ for $j \in \{1, ..., k\}$ in the concept test set, we extract representations $\mathbf{Z}_{c_j}$ at the latent space axis j aligned with concept c_j. These represent the *predicted probability* that image $\mathbf{X}_{c_j}$ is similar to the concept learned at axis j. Based on the ground-truth labels for concept images $\mathbf{X}_{c_j}$, we assign label 1 to all images with ground-truth labels equal to the label of the concept aligned with axis j, and all others label 0. We use these predicted probabilities and binary labels to calculate the one-vs-all AUC score. A score of 1 proves latent representations for concept c_j obtained with exemplary images of concept c_j are perfectly distinguishable from the representations for concept c_j obtained with exemplary images of all other concepts $c_i \in C$ with $i \neq j$.

Concept Representation The different concepts used in the CW method are represented to the explainee (e.g., domain experts) using exemplary images containing the concept of interest according to the corresponding formal definition of the concept. We take these images from the dataset used in the classification task. By showing the concepts to the explainee using examples, the mental model of the explainee is more prepared for receiving image-based explanations of the CW method on the learning and importance of concepts.

4.3 Concerns

Although the CW method focuses strongly on independence between concept representations in the latent space, Mahinpei et al. [54] have shown the provided interpretations can still be misleading due to a wrong design choice that can lead to *information leakage*. More specifically, to measure the purity of concepts, the feature maps that demonstrate the activation of concept $c_j \in C$ for input image $x_1, ..., x_n \in X$ at layer l, defined as the latent space representation of the instance $\mathbf{Z}_{c_j}$, are summarized into an activation scalar. However, since this summarization only occurs during optimization of rotation matrix Q and not in the other training steps nor during testing, the model can *by definition* use more information during training and testing than is used to measure the concept purity. In the worst case, this could lead to a situation where the concept purity results are satisfactory, having the user conclude that the concepts are distinguishable and that the interpretation of the model's decisions in terms of the importance of the learned concepts for predictions are valid. However, this interpretation does not reflect the behavior of the model, leading to unfaithful results. Only when the model has access to the same concept-related information used to measure the purity, the interpretation is truthful to the model's computations. Hence, this is something to take into account when using the concept whitening method.

4.4 Summary

In this chapter, we explained the concept-based interpretability method used to interpret the image-based pre-miRNA detection results. We described it combines the learning of a classification task with that of pre-defined concepts. In this way, the model can explain its classification decisions using the learned concepts. Next, we described which concept-based explanations can be derived from the model. We ended by expressing our concerns about a wrong choice in the design of the method, which may lead to misleading explanations.

Chapter 5

Concept Whitening for Image-based Pre-miRNA Detection

In this chapter, we introduce our framework for interpretable pre-miRNA detection. In Figure 5.1, we provide the components and their relationships. The concept-based interpretability method of interest (i.e., *Concept Whitening* [13]), requires input data and an existing model. This input data consists of data used to solve the prediction task and concept data. For the former, we use the non-coding RNA sequences in the *modhsa*-dataset encoded as images using the algorithm of Cordero et al. [16] (see section 2.3), the latter, we derive from this dataset. The existing model, which should be a pre-trained convolutional neural network, is also taken from Cordero et al. [16], where we apply small transformations to the architecture. With these components, the interpretability method learns to solve the pre-miRNA detection task and provide concept-based explanations for its predictions.

We have described the concept-based interpretability method and encoding algorithm in sections 2.3 and 4.1, respectively. In the next sections, we explain the other components (i.e., the concepts, transformed existing model and concept-based explanations), which we developed in this thesis.

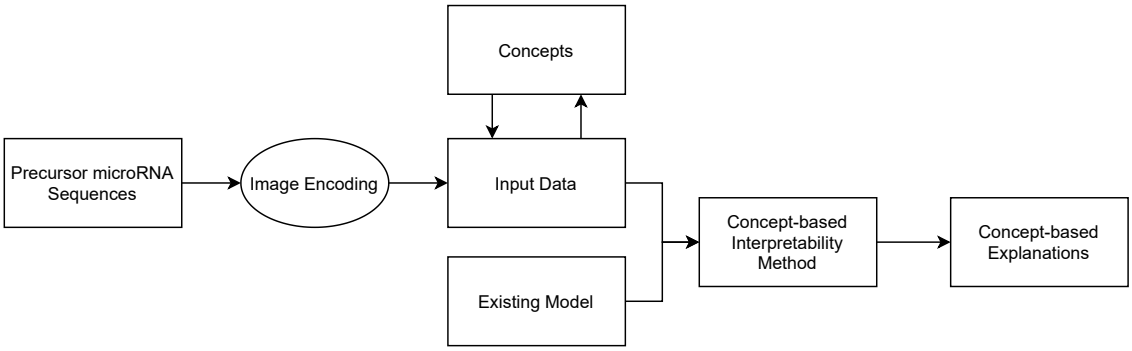

Figure 5.1: The interpretable pre-miRNA detection framework components and their relationships.

5.1 Model Architecture

To apply the CW method, a pre-trained CNN is required. Since Cordero et al. [16] compared the classification performance of several CNNs with different architectures trained on the pre-miRNA detection task, we can one of these CNNs as our pre-trained one. Namely, we choose to apply

the CW technique to their best-performing CNN. Doing this with success requires implementing some adjustments to the CNN. In the following sections, we introduce the architecture of this best-performing model and explain which transformations we apply. Afterwards, we describe the training procedure for this resulting model.

5.1.1 DeepMir

The model architecture with the best performance on the image-based pre-miRNA detection according to Cordero et al. [16] has a VGG-like structure. From Figure 5.2(a) showing the architecture, we find that it consists of three convolutional blocks, followed by a max-pool and a dropout layer. The final block is connected to a fully connected (FC) layer, which receives a flattened version of the output of the convolutional block. Finally, one more dropout and FC layer follow, and a softmax activation function converts their output into class predictions.

5.1.2 Transformations

As explained in section 4.1, applying CW requires transforming a layer l from pre-trained CNN f into a CW layer. Due to a similar normalization procedure used by batch normalization (BN) [33] and CW layers, changing a BN layer into a CW one is likely to yield the best classification accuracy results [13]. The architecture of the original CNN f of DeepMir (Figure 5.2(a)) does not contain any BN layers. Hence, we first ensure it includes at least one and then convert one of the BN layers into a CW one. Moreover, in section 4.3 we explained CW models could provide misleading explanations due to a wrong decision in the CW method's design. Therefore, we apply another transformation to prevent this information leakage issue through the concept feature (or activation) maps by converting them to scalars. Finally, we remove one of the FC layers to increase the interpretability of the network. Next, we explain these transformations in more detail.

Dropout to Batch Normalization As shown in Figure 5.2(a), the DeepMir CNN does not contain any BN layers. Since their goal of preventing overfitting is similar, we replace the three dropout layers in the convolutional base with BN ones, combined with ReLU activation functions. These functions clip negative activation values to 0 to deal with vanishing gradients. In this way, the model can learn more complex relationships otherwise lost [71]. In Figure 5.2(b), we show the model architecture after applying these transformations.

Batch Normalization to Concept Whitening Next, we convert one of the three BN layers into a CW one, depending on the types of concepts in our concept set. Namely, CNNs hierarchically learn a non-linear mapping from input to output using the stacked layers defining the network [70]. The first layers learn low-level information such as textures and colors, while the final ones focus on high-level info such as objects [38]. Consequently, CW's success depends on the location of the CW layer performing the concept alignment, and subsequently, on the concept set. Since our concepts represent objects, we change the last BN layer into a CW one.

Feature Maps to Scalars Then, we transform the architecture to deal with the concerns on (potentially) faulty interpretations [54]. As shown in Figure 5.2(c), we summarize the feature maps obtained from the CW layer into scalars by taking the mean of the max-pooled maps. As a result, the model can only access the same information used to measure the concept purity and importance. Figure 5.3 in section 5.2 illustrates this procedure is graphically.

Removal of Fully Connected Layer Lastly, we remove the first fully connected (FC) layer to facilitate using the model's computations for generating explanations, as shown in Figure 5.2(c). In this way, the scalars from the previous layer, multiplied by the weights of the FC layer generating predictions, explain the importance of the concepts learned in the CW layer for prediction. We base this type of reasoning on Chen et al. [12], who link values representing the similarity between learned prototypes and input images to class weights.

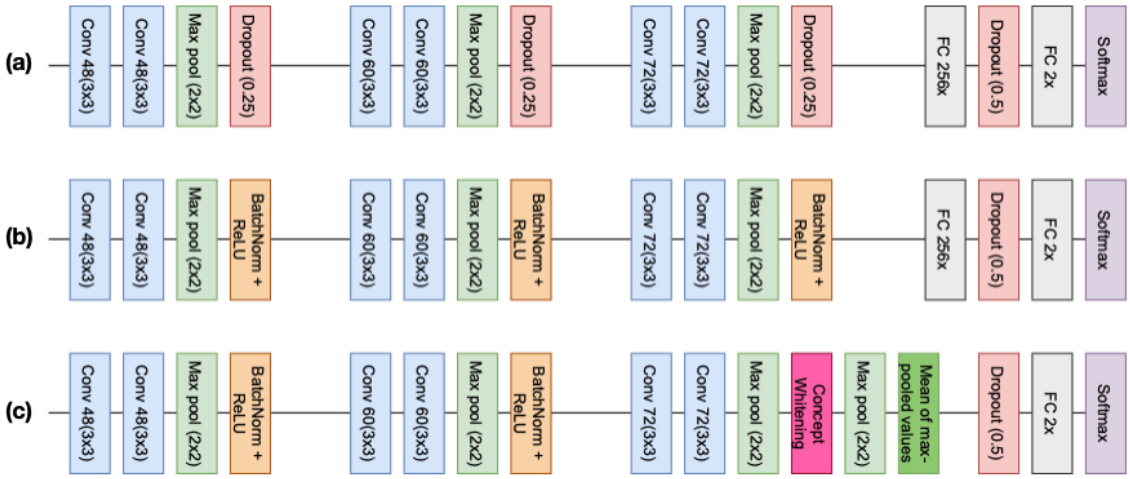

Figure 5.2: Transformations applied to the best-performing CNN of Cordero et al. [16] shown in (a). In (b), we adapt the first three dropout layers into batch normalization (BN) layers with ReLU activations. In (c), we convert the last BN layer to a CW one and replace the first FC layer with two layers computing the mean of the max-pooled feature maps obtained from the CW layer.

5.1.3 Training procedure

Since CW requires an existing (or pre-trained) model, we need weights learned by training a model on the classification task of interest. We cannot use those from the original DeepMir CNN due to the applied transformations. Therefore, before applying CW, we pre-train our adapted CNN on the pre-miRNA detection task. More specifically, we pre-train the model in Figure 5.2(c) before having converted the BN layer to a CW one. We do this similarly to Cordero et al. [16], meaning we first pre-train the model with a large dataset and fine-tune it using the classification dataset of interest. Consequently, we obtain weights we can use to initiate the CNN applying CW in its CW layer (Figure 5.2(c)). Next, we provide a more detailed explanation of the training procedure.

Pre-training First, we pre-train the CNN shown in Figure 5.2(c), given its CW layer is still a BN one, using the *modmirBase*-dataset. The balanced dataset, created by Cordero et al. [16], consists of 24,801 encoded pre-miRNA sequences not in the dataset used in DeepMir [16] to solve the pre-miRNA detection task. We train the model for 20 epochs, using a batch size of 128, Adam-optimizer with a learning rate of $1.0e^{-4}$, and *cross-entropy loss* function.

Fine-tuning Next, we resume the training of the previous model using the *modhsa*-dataset. This dataset, also created by Cordero et al. [16], consists of 3,660 encoded pre-miRNA sequences, with balanced class labels. The model is trained for 100 epochs, using a batch size of 128, the Adam-optimizer with a learning rate of $1.0e^{-4}$, and the *cross-entropy loss* function.

Concept Whitening Model To apply CW, we convert the last BN layer of the previous CNN to a CW one (Figure 5.2(c)) and initiate all other layers with the weights of the best performing model during fine-tuning. The CW model switches between learning the pre-miRNA detection task from the *modhsa*-dataset and performing the CW procedure after every 30 input images. The concept dataset used during the CW procedure contains example images of the six concepts defined in section 5.3 derived from the *modhsa*-dataset. We train this model for 100 epochs with a batch size of 12 and a learning rate of $1.0e^{-3}$, which we multiply by a factor of 0.1 after every 30 epochs.

Besides, two generalization techniques are applied to prevent the model from overfitting the training set. We use 5-fold cross-validation during model training by dividing the original training set of the *modhsa*-dataset, consisting of 2,352 images, into five equally-sized, stratified parts. We

use each part as a validation set once and combine the remaining four into a training set. We create the concept training sets by deriving example images from the training set consisting of four data folds. Similarly for the concept validation set. Consequently, we obtain five different concepts sets for the cross-validation procedure. Besides, an early stopping function terminates the training if the validation loss has not decreased over the last 20 epochs.

5.2 Additional Explanations

With the CNN shown in Figure 5.2(c), different explanations can be given as proposed by Chen et al. [13]. Namely, we can extend the activation-based explanations specifying the importance of concepts for the pre-miRNA class predictions. As explained briefly in section 5.1.2, the summarization of the CW layer's feature maps into scalars enables creating explanations on the importance of concepts for predictions directly from the model's computations. More specifically, for a given input image, each node in the CW layer provides activation values representing the similarity between the image and pattern learned by the node. Since we align some of these nodes with concepts, their activation values explain the similarity between the image and the learned concept representation. Next, in the final layers, the model applies the computations shown in Figure 5.3. The feature (or activation) maps, denoted by squares, are compressed into scalars, denoted by circles. These scalars are linked directly to the nodes in the FC layer, which compute the output values based on a weighted combination of the scalars. Since this layer contains two output nodes, one for each pre-miRNA class, we obtain vectors of weighted activation values for both classes for every input instance. Finally, the model sums the values in both vectors and transforms them into class probabilities. Hence, the model bases the class probability for each output option on the weighted activation of concepts and other information learned in the CW layer.

Since the vectors of weighted activation values also include concept activation values, we can derive the importance of a learned concept for predicting a class label from the vector containing the values of the class of interest. Hence, we define a *concept influence*-metric CI as follows:

$$CI(x_i) = c_{activation}(x_i) * w_{\hat{y}} | f(x_i) = \hat{y} \quad for \quad \hat{y} \in \{positive, negative\} \tag{5.1}$$

with $f(x_i)$ the class prediction made by CNN f for instance $x_i \in X$ for $i \in \{1, ..., n\}$, $c_{activation}(x_i)$ the concept activation scalar of a given concept obtained on image x_i, and $w_{\hat{y}}$ the class weights of the final FC layer. With this measure, a *local* explanation based on our pre-defined concepts and all remaining information can be given. Averaging all local ones enables to provide a *global* explanation that holds for a complete pre-miRNA class.

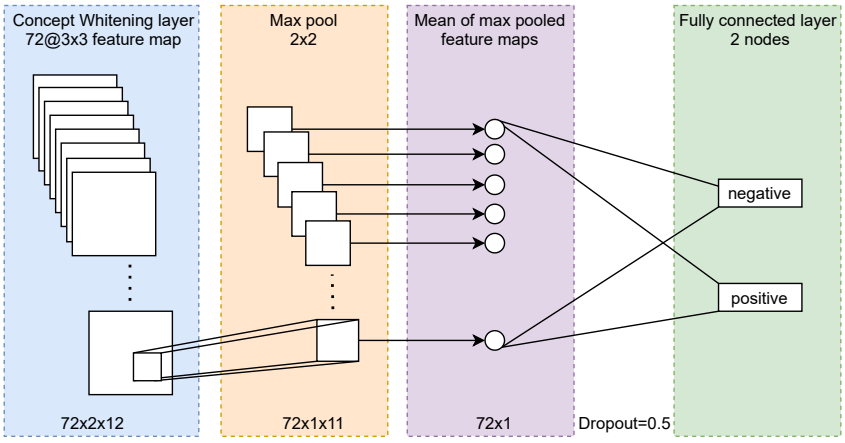

Figure 5.3: Graphical representation of the CW model's computations in its final layers. The model summarizes the feature maps provided by the CW layer into scalars. After weighting and summing these scalars, the model converts them to class probabilities.

5.3 Concepts

As shown in Figure 5.1, two key components required for concept-based interpretability using CW are input data, including concepts, and an existing CNN. In CW, concepts are represented by example images. To the best of our knowledge, we are the first to apply concept-based explanations to the pre-miRNA detection task. Consequently, the first step of our interpretable pre-miRNA detection framework focuses on defining meaningful concepts for this task. For this, we draw inspiration from three sources: (1) previous findings on structural characteristics of (pre-)miRNAs, (2) the encoding algorithm, and (3) saliency results of DeepMir Cordero et al. [16].

Since we collect concept example images from the dataset used for the pre-miRNA classification task, the examples for different concepts can be overlapping in the case of too general definitions. If the concepts are not (close to) independent, we cannot expect CW to learn *independent* representations from their example images. Given this and the fact that Rudin et al. [64] argues defining *good* concepts is an iterative process, we create our concepts as follows. First, we define general concepts and transform the definitions into constraints used to annotate the encoded sequences in the *modhsa*-dataset with the concepts. Afterwards, we compute their presence in both classes of the *modhsa*-dataset and their correlation with other concepts to refine the definitions. Next, we execute the annotation and quantitation steps again, and so on. Finally, we use the most refined concepts for training our CW model. We can use the results from this model and domain expert feedback for further refinement or definition of new concepts.

The refined concepts, the inspiration used to define them, and constraints for annotating the encoded sequences with the concepts are listed below. In appendix B we give examples of more general concepts. The annotation constraints often include the notions width and height, referring to the width and height of image objects in terms of pixels. The total height and width of the images are 25 and 100 pixels, respectively. Besides the definition and annotation constraints, we illustrate the appearance of the concepts in the encoded data. In Figure 5.4(a) we highlight some concepts in an encoded *positive* pre-miRNA, in Figure 5.4(b) this is done for a *negative* instance using the remaining concepts. After providing the concept definitions and annotation constraints, we investigate the specificity and independence of the concepts in the *modhsa*-dataset.

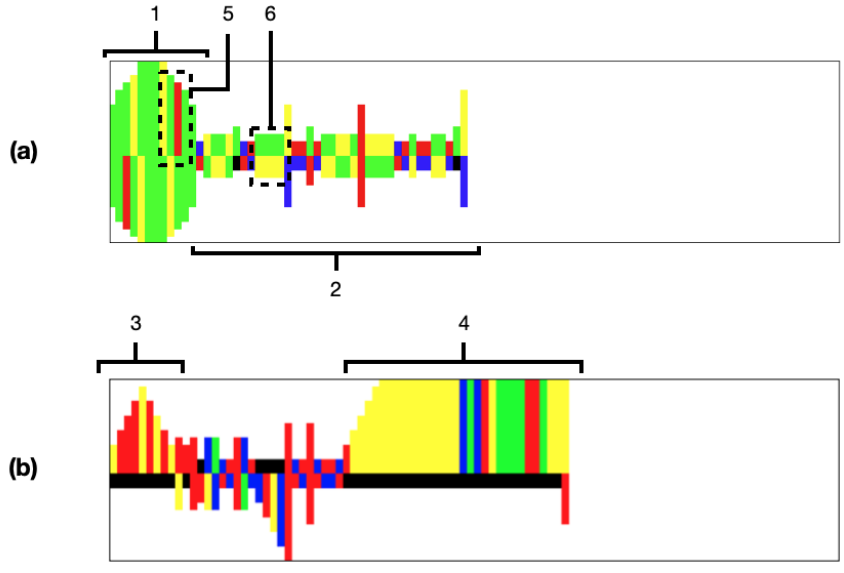

Figure 5.4: Graphical representation of the six pre-miRNA-related concepts. In (a), the *large terminal loop* (1), *at least* 90% *base pairs and wobbles in stem* (2), *U-G-U motif* (3), and *A-U pairs motif* (4) concepts are shown in the encoded *hsa-mir-548ab* pre-miRNA. In (b), the *large asymmetric bulge instead of a terminal loop* (3) and *large asymmetric bulge* (4) concepts are shown in the *hsa_RefSeq_1512* sequence.

5.3.1 Concept Definition and Annotation

Large Terminal Loop Our first concept is the *Large terminal loop*. As explained in the preliminaries on the biogenesis of microRNAs (section 2.2.1), pre-miRNAs have a hairpin-like shape, formed by the terminal loop connecting the 3' and 5' nucleotide strands of the sequence. The enzyme Dicer cleaves this loop to transform a pre-miRNA into a miRNA duplex. Hence, this terminal loop is *by definition* a requirement for all pre-miRNAs. Important to note is that the loop can also occur in other non-coding RNA sequences [2]. We formally define the concept as follows.

Concept 1 ***(Large terminal loop)***

- *Definition: Binary concept specifying whether the sequence contains a terminal loop consisting of at least 12 (potentially) mismatched nucleotides.*
- *Annotation constraints: A terminal loop is recognized in the images if the first sequence pair (going from left to right) does not contain a gap. Two black pixels on top of each other represent a gap. The pixel width and height define the largeness of the loop. More specifically, the width and height should be at least 12 and 21 pixels, respectively.*
- *Example: Figure 5.4(a)1.*

At Least 90% **Base Pairs and Wobbles in Stem** Our second concept *At Least* 90% *Base Pairs and Wobbles in Stem* originates from the finding that the processing of a pre-miRNA into a miRNA duplex by enzyme Dicer is more efficient if the sequence stem contains more base pairs and wobbles [34, 22]. Allmer [2] confirms bulges and loops in the stem should be minimized. In Figure 5.5, Cordero et al. [16] stacked the averages of all their saliency results on top of each other and colored them based on the class prediction. They illustrate the model has a narrow focus on the central stem region for the positive pre-miRNA class, revealing a preference for pairs with short bar lengths (i.e., base pairs or wobbles). The formal concept definition is as follows.

Concept 2 ***(At Least*** 90% ***Base Pairs and Wobbles in Stem)***

- *Definition: A binary concept specifying whether at least* 90% *of the pairs in the pre-miRNAs stem are base pairs and wobbles.*
- *Annotation constraints: For each pair in the sequence's stem, we analyze the color combination and bar heights. Blue-red or red-blue colored bars consisting of 2 pixels each refer to C-G or G-C pairs. Yellow-green or green-yellow colored bars of 3 pixels each define A-U or U-A pairs. Red-green or green-red bars of 4 pixels each represent wobbles (G-U or U-G pairs). To obtain the frequency score, we divide the number of pairs adhering to these constraints by the stem length. We define this length as the number of pixels from the end of the terminal loop until the whitespace area in the right image part. Also, we define the loop end by the first pair after the loop with bar heights less than 3 pixels. We use a threshold of* 0.9 *for the frequency score to only include stems containing at least* 90% *base pairs and wobbles.*
- *Example: Figure 5.4(a)2.*

Large Gap instead of a Terminal Loop We can also define concepts related to the negative pre-miRNA class. As opposite to concept 1, we define the *Large asymmetric bulge instead of a terminal loop* concept. DeepMir's saliency results for the negative pre-miRNA class given in Figure 5.5 highlight the sides of the sequence. Recall that the terminal loop characterizes the left side of positive pre-miRNAs. Other non-coding RNA sequences can have this loop, but it is not required [2]. Hence, this highlighted region may illustrate the absence of the loop is a characteristic of some negative pre-miRNAs. The formal concept definition is as follows.

Concept 3 ***(Large asymmetric bulge instead of a terminal loop)***

- *Definition: Binary concept specifying whether the sequence contains an asymmetric bulge instead of a terminal loop where the bulge consists of at least 10 consecutive mismatches.*
- *Annotation constraints: Given that a terminal loop is recognized if the leftmost sequence pair does not contain a gap, we identify a bulge if this pair does one. We consider the bulge large if its width exceeds 9 pixels.*
- *Example: Figure 5.4(b)3.*

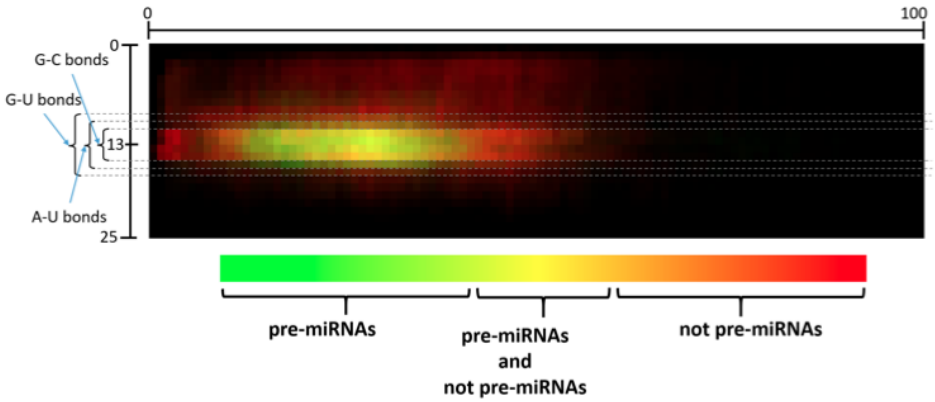

Figure 5.5: Saliency results obtained in DeepMir [16]. The stacked saliency maps, colored based on their class prediction, illustrate the importance of the central stem area for positive pre-miRNAs. For negative ones, the highlighted area is more widespread and towards the sides of the pre-miRNA.

Large Asymmetric Bulge Associated with the efficient pre-miRNA processing with more base pairs and wobbles in the stem, Kang and Friedländer [36] specified that pre-miRNAs should preferably not contain asymmetric bulges sized equal to or larger than the terminal loop. Additionally, the results in Figure 5.5 illustrate the upper sequence region is important for the negative pre-miRNAs, indicating that large (asymmetric) bulges, represented by high bars, are characteristic of negative pre-miRNAs. We formally define the *Large asymmetric bulge* concept as follows.

Concept 4 ***(Large asymmetric bulge)***

- *Definition: Binary concept specifying whether the sequence contains an asymmetric bulge consisting of at least 15 consecutive mismatches.*
- *Annotation constraints: An asymmetric bulge is represented a sequence of consecutive gaps, which are black-colored pixels. We consider the bulge large if its highest bar reaches the image border. Also, the bulge's width should be at least 15 pixels.*
- *Example: Figure B.1(b)4.*

U-G-U Motif Sequence motifs, which are often used in biology to specify nucleotide patterns [30], may also be relevant for concept definition. Nguyen et al. [60] identified that some pre-miRNAs contain a U-G-U motif in the apical stem region of the terminal loop. More specifically, as shown in Figure 5.6 of Nguyen et al. [60], this region is defined as the 5' strand stem region that includes part of the terminal loop and a section before the terminal loop. A formal definition of the *U-G-U motif* concept is the following.

Concept 5 ***(U-G-U motif)***

- *Definition: Binary concept defining whether a sequence of U-G-U or G-U-U nucleotides is present in the terminal loop end or 3 nt after the end of the 5' nucleotides strand.*
- *Annotation constraints: The U-G-U motif is a sequence of red-green-red or green-red-green consecutive pixels in the 5' strand from the middle to 3 nt after the end of the terminal loop. The 5' strand is the nucleotide sequence in the upper half of the image. The highest bars in the loop define its middle, its end is defined by the first base pair or wobble after the loop end.*
- *Example: Figure 5.4(a)5.*

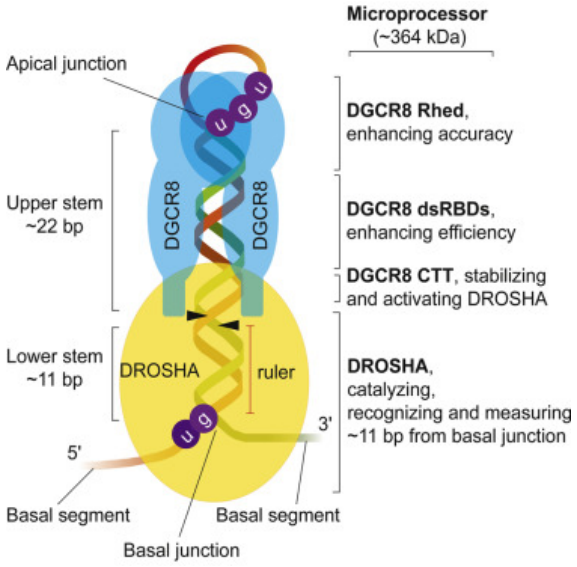

Figure 5.6: Graphical representation of a primary microRNA sequence including the apical U-G-U motif [60]. The motif is located at the end of the terminal loop in the 5' strand.

A-U Pairs Motif Besides the U-G-U motif, Fang and Bartel [22] found that a sequence of A-U or U-A pairs as first pairs of the mature miRNA can stimulate miRNA processing by the enzyme Argonaute. The first miRNA pairs are the pairs following the part of the pre-miRNA cleaved by Dicer. We formally define the *A-U pairs motif* concept as follows.

Concept 6 ***(A-U pairs motif)***

- *Definition: Binary concept specifying whether the sequence contains a subsequence of two or more consecutive A-U or U-A pairs in the 18-25nt region of the mature miRNA.*
- *Annotation constraints: The A-U pairs motif consists of a sequence of consecutive yellow or green pixels combined in a pair such that two bars stacked on top of each other form a combination of green and yellow pixels. Moreover, the bars should each be 3 pixels long. Finally, this sequence should be located in the 18-25 nt region of the mature miRNA. Counting pixels from the first colored pixel on the right to left, we define this region as the pixels on the 18th until the 25th index.*
- *Example: Figure 5.4(a)6.*

5.3.2 Concept Quantitation

After annotating the *modhsa*-dataset images using the concept annotation constraints, we analyze the concepts' specificity and independence.

Concept Presence We evaluate the concepts' specificity based on their presence in the pre-miRNA classes in the *modhsa*-dataset. In Table 5.1 the presence statistics are given. They illustrate the *large terminal loop* (concept 1) and two motifs (concepts 5 and 6) are more present in the positive class. The *at least* 90% *base pairs and wobbles in stem* concept (concept 2) is not present in the negative class, while the *large asymmetric bulge instead of the terminal loop* (concept 3) and *large asymmetric bulge* (concept 4) concepts are only present in that class. Low presences are not an issue since the model looks for similarities between input image parts and learned concepts during prediction. Hence, the images do not need to contain a learned concept completely.

Concept Presence	Positive class	Negative class
Presence of a large terminal loop	12%	4%
Presence of at least 90% base pairs and wobbles in stem	10%	0%
Presence of a large asymmetric bulge instead of terminal loop	0%	6%
Presence of a large asymmetric bulge	0%	7%
Presence of U-G-U motif	26%	9%
Presence of A-U pairs motif	40%	11%

Table 5.1: Presence of the six concepts in the positive and negative class of the *modhsa*-dataset.

Concept Correlation The CW method learns to disentangle concept representations in the CW layer's latent space, requiring (close to) independent (or uncorrelated) concepts. In Figure 5.7 the correlations between the six concepts are given. They show that most concepts are more or less uncorrelated, with correlation values between −0.1 and 0.1. Concepts 2 and 6 are most correlated. Since the motif represents a sequence of base pairs in the stem, this high value is obvious. Other positive correlations originate from overlapping concept images.

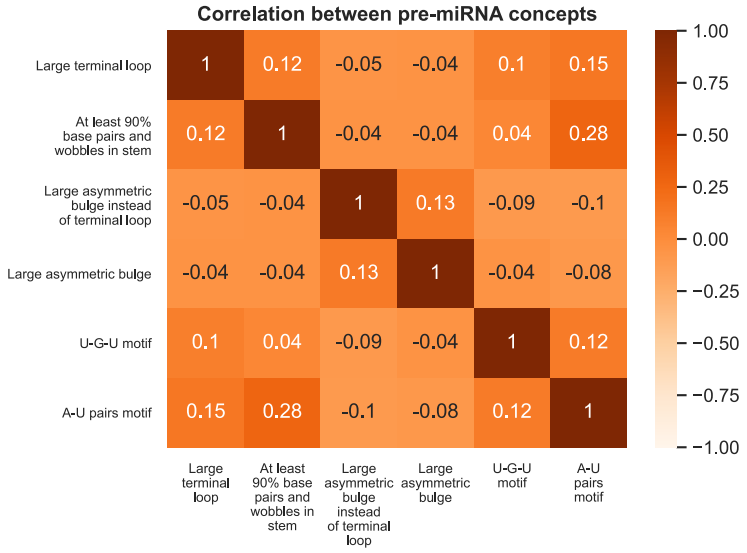

Figure 5.7: Correlations between the six pre-miRNA concepts defined in section 5.3. The darker the orange color, the more positive the correlation. Vice versa, the lighter, the more negative.

5.4 Conclusions

In this chapter, we introduced the framework for interpreting the pre-miRNA predictions using pre-defined concepts. In Figure 5.8 a graphical representation framework is given. The first step is the definition of concepts based on three inspiration sources. Next, we refine the concepts and use them to train our CNN applying the CW technique. From this model, we can generate explanations concerning the learning of concepts and their importance for the positive and negative pre-miRNAs classes. Finally, we evaluate the complete framework and results with a domain expert, whose feedback we can use to improve the concepts and model results [54, 64].

From Figure 5.8, it is clear that the *concept definition, annotation, and quantitation* stage is finished. Therefore, in the following chapter, we head to the green and yellow parts of the framework, where we perform the *concept alignment* and *concept-based interpretability and evaluation* stages.

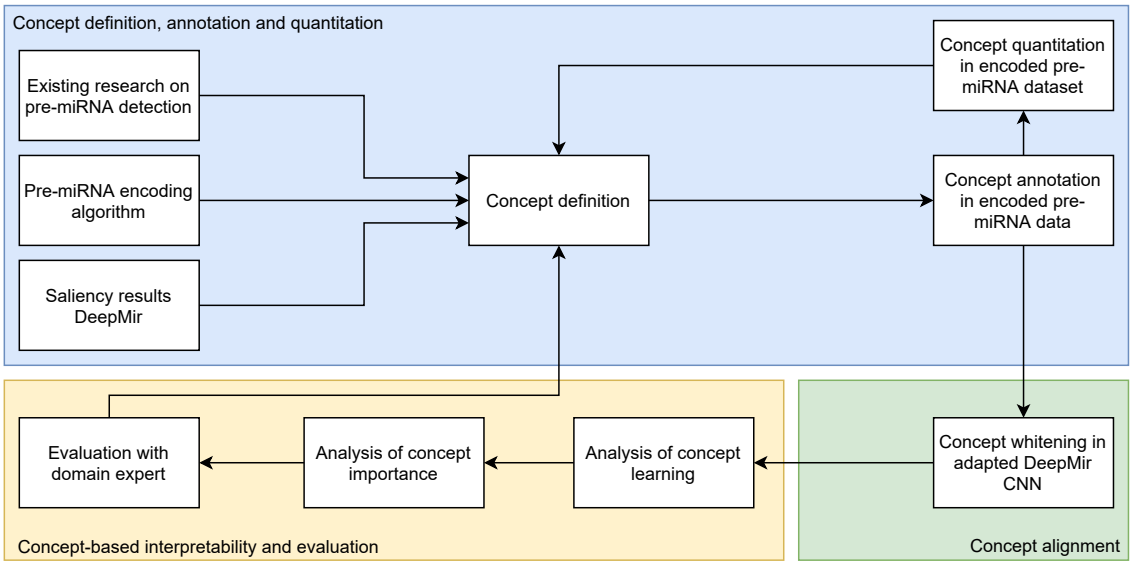

Figure 5.8: The proposed research framework for concept-based interpretability of the pre-miRNA detection results. First, concepts are defined from three inspiration sources. Next, they are refined and used in a CW model. Afterwards, the concept learning and importance information obtained from the resulting model is analyzed. Finally, the results are evaluated with a domain expert.

Chapter 6

Evaluation

In this chapter, we evaluate the results of our proposed concept-based interpretability framework. First, we present the obtained pre-training classification accuracies for the pre-miRNA detection task in section 6.1. In section 6.2, we report accuracy scores for our concept whitening model and interpret its classification decisions using the pre-miRNA-related concepts. In section 6.3, we discuss findings of a domain expert evaluation of the results. We end this chapter with conclusions derived from the results in section 6.4.

6.1 Pre-training

As explained in Chapter 5, applying the concept whitening method requires a pre-trained CNN f, of which one layer is changed into a CW one. Due to their similarity with CW layers, batch normalization (BN) layers are most suited for the change [13]. Since the best performing CNN $f_{original}$ of DeepMir [16] does not contain any BN layers, we replaced its dropout layers with BN layers. Consequently, pre-training the new model f on the pre-miRNA detection task is required.

First, we train f using the pre-miRNA dataset *modmirBase*, consisting of $24,801$ instances with balanced class labels, leading to an accuracy score of $\sim 96\%$. Next, we fine-tune f using the balanced pre-miRNA dataset *modhsa* with $3,660$ instances. In Figure 6.1, the accuracy per epoch obtained on the training and test set is shown. In Figure 6.2, the corresponding loss values per epoch are given. The highest train and test accuracies are $\sim 97\%$ and $\sim 93\%$, respectively. Compared to those obtained with $f_{original}$, we find that both scores dropped by 2%. Thus, the adaptations applied to $f_{original}$ have caused a small performance decrease.

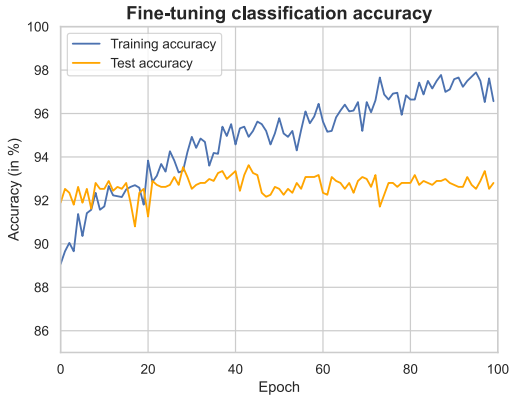

Figure 6.1: Fine-tuning accuracy per epoch obtained on the *modhsa* train and test set.

Figure 6.2: Fine-tuning loss per epoch obtained on the *modhsa* train and test set.

6.2 Concept Whitening Model

To create the CW model, we convert the last BN layer of CNN f to a CW layer and import the pre-training weights in the layers to use the knowledge acquired on the pre-miRNA detection task. Below, we discuss the model's obtained accuracy scores and concept-based interpretability results. The latter illustrate the model's ability to learn our pre-miRNA concepts and their importance for the classification task. As discussed in section 4.1, their representations in the CW layer should be *pure* to provide faithful explanations. Therefore, we first discuss the concept purity results, followed by the quality of the learned concepts and their importance for the detection task. Finally, we try to derive new concepts from the CW's residual information handling.

6.2.1 Concept Whitening Model including all Pre-miRNA Concepts

We train the concept whitening CNN f using all six pre-miRNA concepts defined in section 5.3 by aligning each of them with a different node of the total 72 nodes in the CW layer. The remaining 66 nodes handle residual information not included in the concept example images.

Pre-miRNA Classification Performance Since we train CNN f using 5-fold cross-validation, we average the classification scores obtained on the five folds and report this average and associated standard deviation. Using the weights of the model with the highest fine-tuning test accuracy, we obtain an average training and validation score of 73.32% (±2.51) and 87.0% (±6.66), respectively. In Figure 6.3, we show the training accuracy per epoch for all five folds and in Figure 6.4 the accuracy per epoch obtain on the validation folds. We notice the training accuracies drop around epoch 25, while only the validation score obtained on the fourth validation fold decreases rapidly after 35 epochs. The other validation scores remain high since the early stopping function monitoring the decrease in validation loss terminates training. The accuracy drops likely indicate that the disentanglement of the CW layer's latent space takes effect around epoch 25. Using higher learning rates instead of the current $(1.0e-3)$ to enforce more rigorous disentanglement learning leads to scores stagnating around 50.0%. From this, we conclude that f cannot successfully disentangle the CW layer's latent space without failing to learn the underlying classification task.

We evaluate the performance on the *modhsa* test set using the model with the highest validation accuracy, which is obtained using the second validation fold. The corresponding model achieves a test accuracy score of 91.26%, meaning 999 of the 1,098 instances are correctly classified. The score is much higher than the training scores. However, removing the early stopping to allow for longer training would likely illustrate a decrease in test accuracy similar to the decreasing training accuracies.

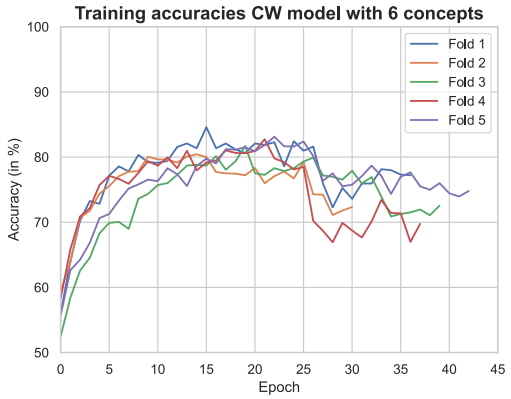

Figure 6.3: Accuracy per epoch using the CW model, all pre-miRNA concepts and the five training *modhsa*-data folds.

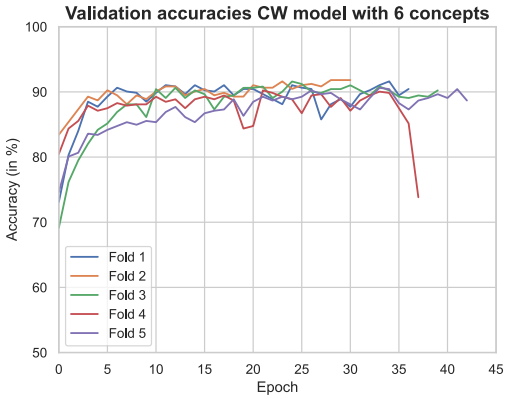

Figure 6.4: Accuracy per epoch using the CW model, all pre-miRNA concepts and the five validation *modhsa*-data folds.

Concept Purity As mentioned in the introduction of this section, the concept learning and importance results depend on the disentanglement of the CW layer's latent space that includes the learned concept representations. Therefore, we first investigate the quality of the disentanglement by analyzing the purity of these representations using two different approaches.

The first approach is based on the inter-concept similarity metric 4.3. In Figure 6.5, the normalized inter- and intra-concept similarity scores for the representations of the six concepts in the CW layer's latent space are given. The darker orange the color, the higher the similarity. This figure shows that the model finds several representations of different concepts highly similar. For instance, the similarity between concepts *Large asymmetric bulge* (concept 4) and *Large asymmetric bulge instead of terminal loop* (concept 3) is 0.9, meaning they are almost identical. This is similar for other concept combinations, such as the two motifs. The similarity between representations of concepts *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge instead of terminal loop* (concept 3) and is much lower, namely 0.42. Hence, they are more separable. This is similar for the representations of concept 2 and several others such as concepts 1 or 4.

The second approach for evaluating the concept purity uses concept AUC scores. They illustrate the model's ability to predict concepts from representations in the CW layer. These representations are activation values of concept example images at the nodes in the CW layer aligned with concepts. Similar to Chen et al. [13], we calculate these one-vs-all AUC scores by splitting the test set of concept example images into five equally-sized parts. In Figure 6.6, the averages and standard deviations of the five AUC scores for all six pre-miRNA concepts are given. Concepts *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) have average AUC scores of 0.79 and 0.90, respectively. These scores indicate that the model can distinguish their representations well from those of the other concepts. The remaining scores are 0.6 or lower, illustrating the representations of these concepts are difficult to distinguish from others.

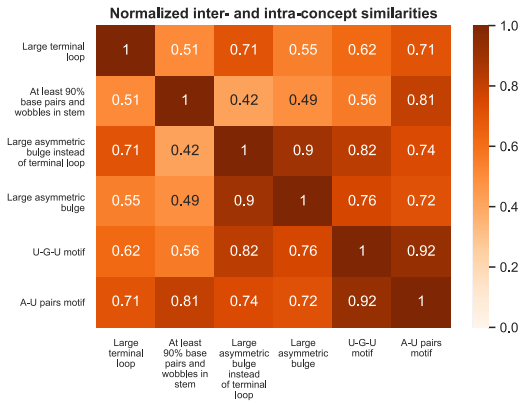

Figure 6.5: Normalized inter- and intra-concept similarities between concept representations in the CW layer's latent space aligned with the six pre-miRNA concepts.

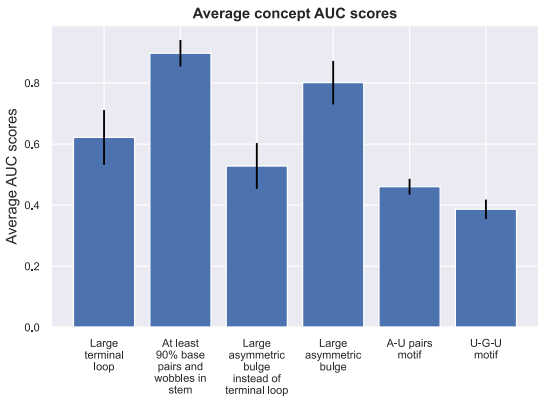

Figure 6.6: Average one-vs-all AUC scores calculated using the activation values of concept test images in the CW layer aligned with the six pre-miRNA concepts.

Conclusions From the concept purity results (Figure 6.5 and 6.6), we find that most learned representations of different concepts are too similar to be disentangled by the model. From the concept correlation values given in Figure 5.7, we can derive that even small *positive* correlations between concepts already demonstrate the model cannot disentangle their representations in its latent space. It may also be that the example images of different concepts look too similar. As explained in section 4.2, issues with the purity of concepts can result in misleading concept-based explanations. Therefore, we conclude that the model aligned with all six pre-miRNA concepts is unsuitable for interpreting the concept learning and importance for the pre-miRNA class predictions.

6.2.2 Concept Whitening Model including the `At least` 90% `base pairs and wobbles in stem` (concept 2) and the `Large asymmetric bulge` (concept 4) Concepts

In the previous section, we concluded that using all six pre-miRNA concepts in the CW model gives undesirable results. As a solution, we could use only the concepts that seem to have (close to) pure representations in the latent space. From Figure 6.5 and 6.6, we find this seems to be the case for concepts 4 and 2. Therefore, we train a new CW model that learns these two concepts in the first two nodes of the CW layer. All remaining 70 nodes in this layer search for features useful for the pre-miRNA detection task in the residual image information.

Pre-miRNA Classification Performance We train the CW model aligned with the two concepts using 5-fold cross-validation and obtain a training and validation accuracy averaged over all folds of 92.40% (±1.94) and 92.70% (±1.13), respectively. In Figure 6.7 the training accuracy per epoch for the five data folds is given. The figure shows the model obtains scores above 90% on all folds. The model trained with the second data fold finished all 100 epochs, indicating the validation loss kept improving. For the other folds, the early stopping function terminated training before reaching the maximum number of epochs. Figure 6.8, which shows the accuracy scores per epoch obtained on the validation sets, demonstrates all validation accuracies lie between 90% and 95% and the model trained on the second fold scores best. Compared to the results obtained with the model aligned with all six pre-miRNA concepts (section 6.2.1), we find no drops in accuracy, meaning the training is more stable. Hence, it appears that this model is better at simultaneously learning the pre-miRNA detection task and disentangling the CW layer's latent space.

We evaluate the performance on the test set of the *modhsa*-dataset using the model with the highest validation accuracy, which is the one trained on the second data fold. The corresponding test accuracy of 92.81% is similar to the scores obtained during training and validating. Comparing this score to the one of $\sim$ 95.0% obtained using the original DeepMir CNN [16], we can conclude that the CW model has given in only 2% on the classification performance. Given the CW model should be interpretable following the concept alignment, this accuracy drop is very acceptable.

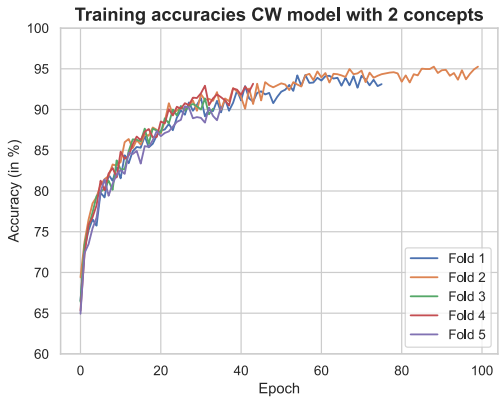

Figure 6.7: Accuracy per epoch using the CW model, *Large asymmetric bulge* and *At least* 90% *base pairs and wobbles in stem* concepts and five training *modhsa*-data folds.

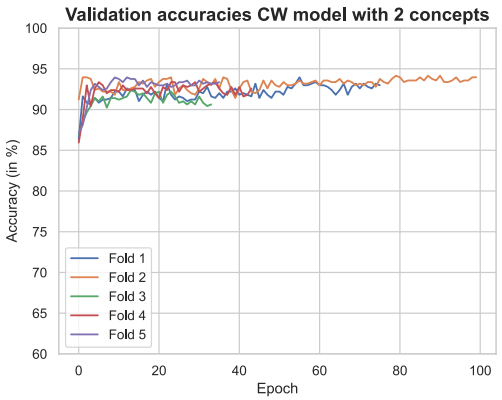

Figure 6.8: Accuracy per epoch using the CW model, *Large asymmetric bulge* and *At least* 90% *base pairs and wobbles in stem* concepts and five validation *modhsa*-data folds.

Concept Purity After evaluating the classification performance, we analyze the purity of the learned concept representations. In Figure 6.9, the normalized inter- and intra-concept similarities between the representations of concepts 4 and 2 in the CW layer are given. The inter-similarity score is 0.35, illustrating the model can separate the representations of the concepts quite well. The average concept AUC scores, shown in Figure 6.10, both exceed 0.85. This score shows that concept activation values obtained with examples of the concept of interest are ranked higher by the model in more than 85% of cases than the values obtained with examples of the other concept. This further illustrates the model can distinguish the two concepts.

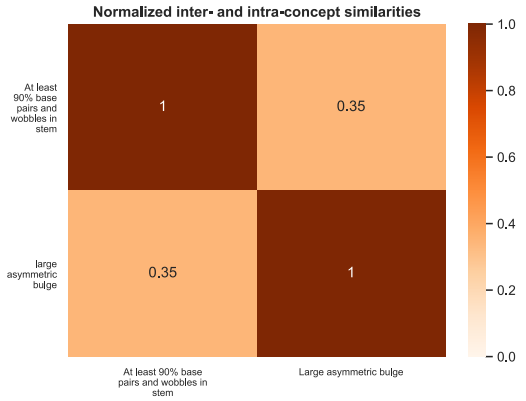

Figure 6.9: Normalized inter- and intra-concept similarities between concept representations in the CW layer.

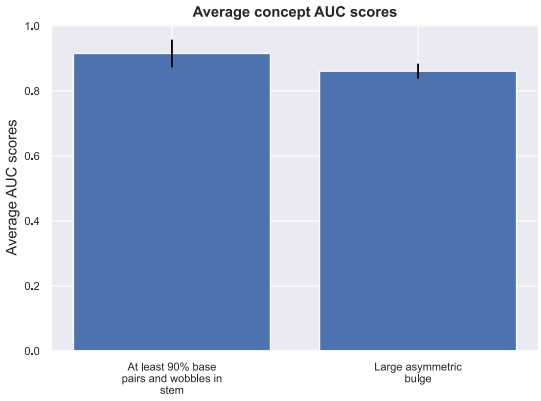

Figure 6.10: Average one-vs-all AUC scores calculated using the activation values of concept test images in the CW layer.

Concept Learning From section 6.2.2, we conclude the model can learn *pure* representations for our two concepts. Consequently, we evaluate the model's representations learned from their example images by visualizing the test images most activated on the latent space node aligned with the concept. We also include the least and moderately activated images to illustrate which images the model considers least or less concept representative. The most activated images include the *normalized receptive field* the concept of interest. As described in 4.2, these fields clarify the focus of the CNN's filter representing the concept. Namely, for each image, we calculate the field values of the two concepts and use them for normalization to scale the values obtained for the two concepts onto the same range. Next, we set all normalized values exceeding the 93^{th} percentile to 1 and all others to 0. In this way, the image regions that are *highly* similar to the learned concept are emphasized.

First, we show the concept learning results for concept 2 and then for concept 4. In appendix C.1, more concept learning results of the model are provided.

- Concept 2: *At least* 90% *base pairs and wobbles in stem*

  The three most activated images and the empirical receptive field of concept 2 are provided in Figure 6.11. All images contain a high number of base pairs, similar to concept 2. Moreover, the receptive field illustrates the model's focus on the image region including only base pairs and wobbles. Figure 6.12 shows the same three images but including the receptive field of concept 4. The fields show the model does not find any region highly similar to the learned concept representation.

  In Figure 6.13 the moderate activated images are shown, whereas Figure 6.14 shows the least activated ones. The moderate activated images contain more base pairs and wobbles in the stem compared to the least activated ones. The latter mostly contain (asymmetric) bulges. Hence, it seems the model has learned a representation for concept 2 in line with our definition of the concept.

Figure 6.11: Top three most activated images for concept *At least* 90% *base pairs and wobbles in stem* (concept 2) including its receptive field.

Figure 6.12: Top three most activated test images for concept *At least* 90% *base pairs and wobbles in stem* (concept 2) including the receptive field of concept 4.

Figure 6.13: Test images with moderate activation values for concept *At least* 90% *base pairs and wobbles in stem* (concept 2).

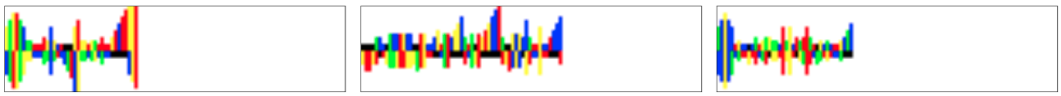

Figure 6.14: Test images with the lowest activation values for concept *At least* 90% *base pairs and wobbles in stem* (concept 2).

- Concept 4: *Large asymmetric bulge*

  In Figure 6.15 the three most activated test images and receptive field the concept 4 are given. All include large asymmetric bulges or several smaller ones. The receptive fields demonstrate the node aligned with concept 4 focuses on the lower right image corner. The model finds colored pixels in this area, combined with black-colored ones to the left or whitespace to the right, concept representative. The black-colored pixels represent gaps, which are part of (large) asymmetric bulges. Long bars, represented by the colored pixels in the lower image region, are also included in large asymmetric bulges. The same three images are shown in Figure 6.16, but with the receptive field of concept 2. They show that the images contain some information highly similar to the model's concept representation. However, this focus changes in every image, complicating the process of understanding which image part the model indicates as most important.

  The moderate activated images are given in Figure 6.17, the least activated ones in Figure 6.18. All moderate activated images contain at least one asymmetric bulge. Comparing them to the most activated ones (Figure 6.15), we find the number of asymmetric bulges lower and smaller in size. The least activated include only some small asymmetric bulges or symmetric ones. From these images, we conclude the learned representations for concept 4 are more or less as expected.

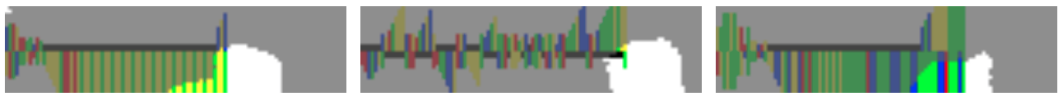

Figure 6.15: Top three most activated images for concept *Large asymmetric bulge* (concept 4) including its receptive field.

Figure 6.16: Top three most activated images for concept *Large asymmetric bulge* (concept 4) including the receptive field of concept 2.

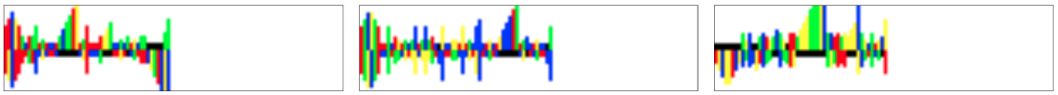

Figure 6.17: Test images with moderate activation values for concept *Large asymmetric bulge* (concept 4).

Figure 6.18: Test images with the lowest activation values for concept *Large asymmetric bulge* (concept 4).

Concept Importance From the concept learning results in section 6.2.2, we have derived the model's concept representations. Next, we define their importance for the *negative* and *positive* pre-miRNA class predictions using the *concept influence*-metric 5.1. This metric determines a concept's influence on the predictions based on the model's computations used to generate the predictions. Since our model contains two output nodes, one for each pre-miRNA class, we obtain two output values for every input instance. We consider them as *evidence* the instance belongs to the class associated with the value. The most positive evidence value defines the class prediction.

We can generate local and global concept influence explanations. Using several example explanations, we first demonstrate the local and then the global variant. Concept 2 and 4 in the local explanations represent concepts *At least* 90% *base pairs and wobbles in stem* and *Large asymmetric bulge*, respectively. In appendix C.2 more explanations of both scopes are given.

- *Local explanations.* Examples of local explanations are as follows.

  1. Consider the encoded sequence, evidence for both pre-miRNA classes, and corresponding interpretation given in Figure 6.19. The sequence contains several base pairs and wobbles in its stem and no large asymmetric bulge. We see concept 4 has no influence on the evidence for the *negative* class. Concept 2 has a negative influence, illustrating the model links concept 2 to the *positive* class, and the image's similarity with the learned concept representation decreases the evidence for the *negative* class. The reverse holds for the *positive* pre-miRNA class' evidence. The remaining 70 CW layer nodes have a large negative influence on the *negative* class and a large positive influence on the *positive* one, illustrating that several nodes learned patterns in favor of the latter class.
  2. In Figure 6.20, we provide another input sequence, class evidence and corresponding interpretation. The sequence contains a large asymmetric bulge and a few base pairs and wobbles. We find concept 4 has a small positive influence for both pre-miRNA classes. Concept 2 has a negative influence on the *negative* pre-miRNA class. Since the model considers some image regions similar to this concept's learned representation, the chance of being a *negative* pre-miRNA decreases. The large positive influence of the remaining nodes makes this negative influence negligible. The large negative influence of these nodes for the *positive* class leads to a *negative* pre-miRNA class prediction.

3. The sequences in the previous two explanations clearly contain one of two concepts of interest, making the provided explanations quite obvious. Therefore, we include an explanation for a sequence less clearly linked to either one of the concepts. In Figure 6.21, we provide the classification evidence for the sequence. The sequence contains some base pairs and wobbles, some small asymmetric bulges, and several symmetric ones. We find the model's class prediction is mostly generated based on the influence of the remaining 70 CW layer nodes. Especially concept 4 has little influence compared to the remaining nodes. Hence, the concept importance explanations are rather limited for sequences not containing image regions similar to the learned concept representations. This stresses the urge to analyze the information learned in the remaining 70 nodes.

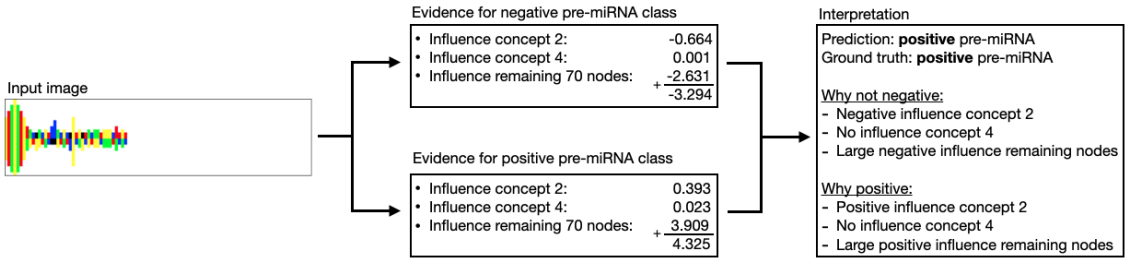

Figure 6.19: Local concept influence explanation for sequence *hsa-mir-148a* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

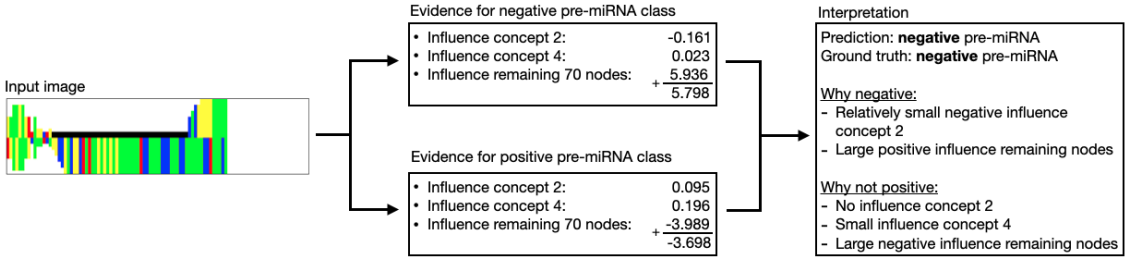

Figure 6.20: Local concept influence explanation for sequence *hsa2_2157* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

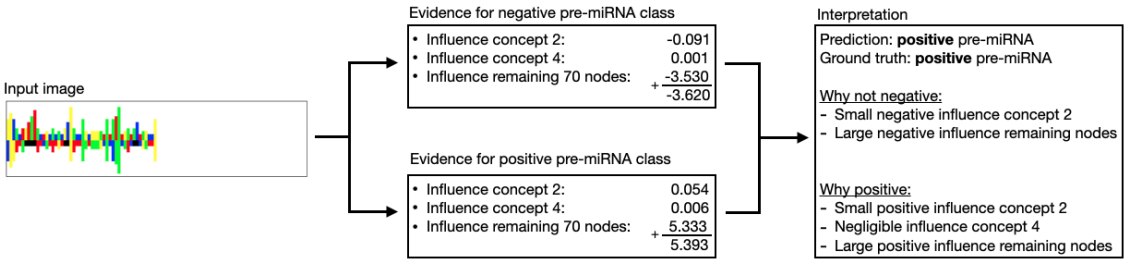

Figure 6.21: Local concept influence explanation for sequence *hsa-mir-30c-1* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

- *Global explanation.* On a global level, we explain the concept influence as follows.
  - *Positive pre-miRNA class.* Consider Figure 6.22, showing the influence of the two concepts, the average influence of all nodes in the CW layer, and the minimum and maximum influence of a single node in the layer. Note that all influence values are averaged over the complete test set. They show that concept 4 has little influence on the positively classified instances. Concept 2 has an above average, positive influence, meaning the model determines this concept relevant for the *positive* pre-miRNA class. Besides, we notice that the most influential node has much more influence on the predictions than our concepts, and there exists a node with a large negative influence on the predictions.
  - *Negative pre-miRNA class.* In Figure 6.23, the average influence values are given for the negative pre-miRNA class. Interestingly, learned concept 4 has no influence on the predictions. Concept 2 has a negative influence, meaning images containing regions similar to this learned concept decrease the evidence for the *negative* class. Similar to the positively classified pre-miRNAs, our concepts have a much lower influence than the most influential node, and one node in the CW layer has a largely negative influence.

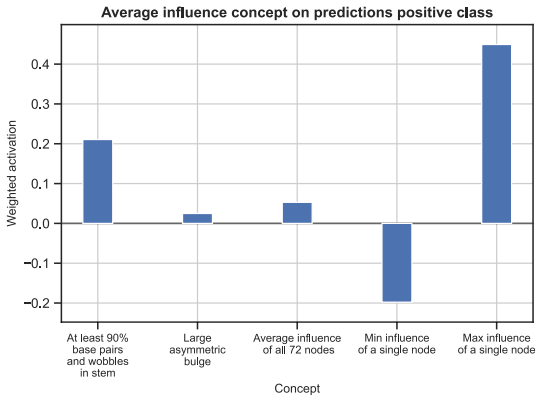

Figure 6.22: Average concept influence value, the average influence of all nodes, and the minimum and maximum average influence of a single node in the CW layer for the *positive* pre-miRNA class.

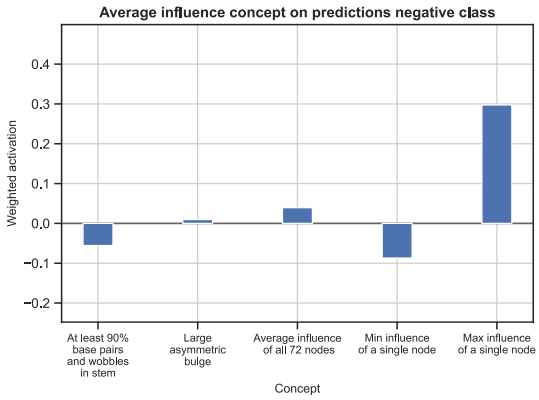

Figure 6.23: Average concept influence value, the average influence of all nodes, and the minimum and maximum average influence of a single node in the CW layer for the *negative* pre-miRNA class.

Definition of New Pre-miRNA-related Concepts Finally, we try defining new concepts from the model's residual image information handling by analyzing shared information among the most activated images for the CW layer nodes not aligned with a concept. We do this only for the most influential nodes since they contain information relevant for the pre-miRNA detection task. In Figures 6.22 and 6.23, the rightmost bars show the average influence of these nodes for the positive and negative class, being nodes 36 and 4, respectively. Below, we explain their learned information. In appendix C.4, we provide more images related to these highly influential nodes.

- *Highly influential node positive pre-miRNA class.* In Figure 6.24, the three most activated test set images for node 36 are given. They demonstrate that the model searches for a combination of green-yellow pixels in the stem area. Interestingly, apart from the location in the pre-miRNA stem not being that important, this pattern overlaps to some extent with our definition of the *A-U pairs motif* concept (concept 6). Note that concept 2 also includes

this green-yellow stem pairs pattern. Comparing the results of the learning of that concept (section 6.2.2) with the results shown here, we can conclude that the focus of node 36 is more on the green-yellow pixel pairs. The node aligned with concept 2 seems to focus more on both green-yellow and red-blue pixel pairs.

- *Highly influential node negative pre-miRNA class.* The most activated images for the fourth node of the CW layer, given in Figure 6.25, contain at least one large asymmetric bulge. Comparing this to the highly activated images for concept 4, we see that node 4 focuses more on the central part of the pre-miRNAs stem in combination with whitespace in the right image region. However, given that the highlighting in all images is not very similar, it is complicated to derive which pattern the model has learned in node 4.

Figure 6.24: Top three most activated images for the most influential node in the CW layer for the *positive* pre-miRNA class (i.e., node 36).

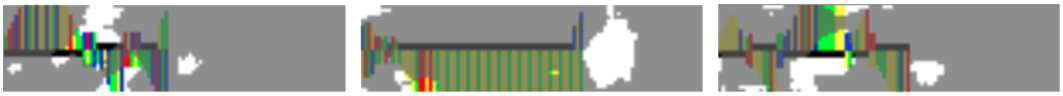

Figure 6.25: Top three most activated images for the most influential node in the CW layer for the *negative* pre-miRNA class (i.e., node 4).

Conclusions From the CW model aligned with concepts 2 and 4, we can conclude the following. The model can successfully disentangle the CW layer's latent space such that the pre-miRNA classification performance is similar to that obtained with a CNN not applying CW. We derived the model's representation of the two concepts based on the most activated test images for the nodes in the CW layer aligned with the concepts. For concept 2 the representations are in line with expectations, for concept 4 this is more complicated to evaluate. Moreover, since the concept representations seem (close to) pure, their importance for the class predictions can be accurately determined. We found only concept 2 to be influential for the class predictions, with a positive influence on the *positive* pre-miRNA class and a negative on the *negative* class. This is in line with existing domain knowledge on pre-miRNA characteristics. However, important to note is that the concept is only influential if the model finds regions of the encoded input sequence that are similar to the learned concept representation. Finally, we evaluated the pattern learned by the most influential nodes in the CW layer which were not aligned with a concept. One of these nodes seems to validate another previous pre-miRNA detection finding, namely the A-U pairs motif (concept 6), except that the location is not as strict as in the definition of the concept.

6.3 Evaluation with Domain Expert

After evaluating the *concept learning* and *concept importance* steps of the framework, we move to the final step for evaluating the concept-based explanations for the pre-miRNA class predictions, namely the *domain expert evaluation*. Evaluations of interpretability methods indirectly assess whether the motivation for interpretability is met [14]. Recall that our motivation is the *promotion of knowledge discovery*. Moreover, as explained in section 3.4, the intrinsic *fidelity* of concept whitening enables to focus on the method's *understandability*. Hence, we evaluate whether the framework and its results provide a domain expert with an understanding of the predictions made by the interpretable pre-miRNA detection model. Furthermore, we analyze whether they can assist

in finding discriminative pre-miRNA characteristics. We combine these two objectives in the notion of *usefulness* and set this as the main focus of the evaluation with a domain expert.

We evaluate the framework by interviewing a domain expert, as this seems appropriate given the fairly preliminary status of the framework and results. This domain expert has contributed to the pre-miRNA image encoding algorithm explained in section 2.3. Furthermore, before the interview took place, we provided the expert with the proposed framework and results. Consequently, before the interview of approximately one hour started, we answered the domain expert's questions on this provided information. During the interview, we focused on four framework aspects: the provided *interpretation* (1), the generation of *concepts* (2), the concept whitening and explanation generation *approach* (3) and the framework in *general* (4). Since we focus on the *understandability* of the framework and its results, we composed the interview questions using the characteristics of this interpretability goal (section 3.2). In appendix C.5 all questions are listed.

(1) Interpretability First, we evaluate the provided interpretability on a higher level. The domain expert immediately pointed out the explanations are difficult to understand due to the distinction between the two pre-miRNA classes when explaining the influence of concepts. Hence, the chosen explanation representation complicates deriving conclusions from the explanations. The domain expert would prefer explanations without this distinction.

Next, we focus on the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts used in our final CW model. The domain expert found the model's learned representation of concept 2 in line with expectations to a great extent. However, the difference in the highlighted areas in Figure 6.11 is confusing. The positive influence of the concept on the positively classified pre-miRNAs follows the current biological knowledge on structural pre-miRNA characteristics. Also, the domain expert found the concept's definition similar to his mental model of the concept originating from the biological knowledge on pre-miRNAs. For concept 4, the domain expert agreed that it is complicated to derive meanings from the learned representations shown in Figure 6.15. More specifically, the large emphasis on whitespace is confusing to understand from a biological point of view.

In terms of the representations of the explanations, the concept learning (section 6.2.2) first did not include the least and some moderately activated images. The domain expert emphasized that adding these images would give a completer view of the learning. Consequently, we have included these images in Figure 6.13, 6.14, 6.17, 6.18 to increase the understandability of these explanations.

Finally, we discussed the explanations concerning the residual information handling in the CW layer. The domain expert found this feature interesting since biologists working on the pre-miRNA detection task are curious about which concepts (or features) the ML model considers relevant for the detection task. They could advance their knowledge of discriminative pre-miRNA characteristics. Regarding the explanations' representations, the focus on the whitespace is again confusing, decreasing the understandability of the explanations. Also, the differing highlighted regions in the most activated images for the node highly influential for the *negative* pre-miRNA class shown in Figure 6.25 complicate deriving meanings from them.

(2) Concepts As explained in section 5.3.1, we defined concepts by leveraging existing pre-miRNA detection knowledge. We represent them using an encoded sequence of the *modhsa*-dataset containing the concept of interest, as shown in Figure 5.4. The domain expert found this representation informative and understandable. Important to note is that the expert has contributed to the pre-miRNA encoding algorithm (section 2.3). Hence, others who have not seen the encoded pre-miRNAs may face difficulties with understanding the representations. Consequently, this would probably require translating the representations to their meaning used in (pre-)miRNA research.

Besides evaluating existing concepts, the domain expert proposed several new ones potentially relevant for improving the results. For instance, the presence of many strong bonds (G-C or C-G pairs) in the middle of the sequence's stem could be a concept related to the *negative* pre-miRNA class. For the *positive* class, weaker pairs such as A-U pairs and wobbles are preferred. Adding this concept could potentially help to explain the model's class predictions to a fuller extent.

(3) Approach Next, we discuss the domain expert's opinion on the information provided for explaining the concept whitening technique itself and the creation of the concept importance explanations. The domain expert could acquire a high-level understanding of the technique and its goal from the provided information. Regarding the concept importance explanations, the domain expert found the use of concept activation values to determine the importance values for the class predictions clear. However, as explained in section 6.3, the distinction between the two pre-miRNA classes complicates understanding the explanations quite severely. We expect the explanations to be more understandable if they only express the reason for a particular pre-miRNA class prediction based on the learned concepts instead of providing evidence for both classes separately and combining this to explain the resulting class prediction. However, this type of explanation is currently challenging or even impossible to generate due to the use of two output nodes (e.g., one for each pre-miRNA class) in the output layer of our CW model.

(4) General Finally, we evaluate the proposed framework for interpreting the image-based pre-miRNA detection results as a whole. Overall, the domain expert found the framework to be potentially useful for the interpretation of the results. However, biologists may likely prefer deriving concepts (or features) on the structural pre-miRNA characteristics from the results and validate them in lab experiments, rather than defining the concepts *beforehand*. The framework aims to combine these to workflows by enabling users to use pre-defined concepts and derive new ones from the residual information handling in the CW layer. However, before evaluating the value of this framework component more extensively, the understandability of the representations of the concept-based explanations should be increased.

Conclusions Overall, the domain expert found the proposed framework interesting. The interpretability method's ability to combine information learned from pre-defined concepts with residual image information (from which new concepts may be derived) to explain the classification predictions is considered valuable. However, the distinction between the two pre-miRNA classes in the local concept importance explanations negatively impacts their understandability, decreasing the overall usefulness of the proposed framework.

6.4 Conclusions

In this chapter, we presented and evaluated the results collected using our interpretable pre-miRNA detection framework. Recall that our chosen concept-based interpretability method (i.e., concept whitening [13]) should simultaneously learn the pre-miRNA detection task and our pre-miRNA-related concepts to explain the class predictions using the learned concepts. Consequently, we first analyzed whether the method can learn independent representations from the example images of our pre-miRNA-related concepts while still achieving high accuracy on the pre-miRNA detection task. This seems to be the case when using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts. The learned representations for concept 2 are in line with expectations, while those for concept 4 are more difficult to evaluate. For the remaining four pre-miRNA-related concepts defined in section 5.3.1, the method seems unable to learn independent representations and obtain accurate pre-miRNA classification performance.

Next, we analyzed the importance of concepts 2 and 4 for the pre-miRNA classification task. We found that concept 2 has a positive influence on the *positive* pre-miRNA class and vice versa for the *negative* class. Recall that this distinction between the two pre-miRNA classes originates from using two output nodes in our CW model, one for each class. Concept 4 seems to have little to no influence on both classes. Note, however, that these conclusions only hold for the input sequences containing image regions similar to the learned representations of these concepts.

Finally, a domain expert evaluation illustrated the framework's potential practical usefulness for interpreting the pre-miRNA detection results. Several aspects should be improved to increase the overall understandability, most important of which is the representation of the concept importance explanations.

Chapter 7

Discussion

In this chapter, we discuss the different aspects of the proposed concept-based interpretability framework. We start with a discussion on our pre-defining concepts for concept-based interpretability methods in section 7.1. In section 7.2, we consider the different results provided by our concept whitening model. Afterwards, we briefly discuss our domain expert evaluation.

7.1 Pre-defined Concepts

As mentioned in section 5.3, it can be difficult to create a concept set consisting of *good* concepts that provide useful concept-based explanations for a ML model's predictions [64]. More specifically, defining the concepts' usefulness *before* training and evaluating the model is challenging. After defining our pre-miRNA concepts, we refined the definitions based on the concepts' presence and correlation with others in the two pre-miRNA classes of the *modhsa*-dataset. We assumed this increased the concepts' chance of being discriminative and their representations in the CW model's latent space to be independent. However, the results in section 6.2.1 illustrated most are still too similar, or their example images contain too many similarities to obtain independent representations. Hence, it seems using basic statistics alone to define *good* concepts is suboptimal.

Besides, human annotators usually create a set of pre-defined concepts. They use assumptions about relevant data attributes to define concepts for explaining the predictions made based on the data. However, as Ghorbani et al. [24] explains, this likely introduces human bias in the process, limiting the explanatory power of the concept-based explanations since the most relevant features may be missed. For instance, it can be challenging for humans to define concepts specifying the relationship between multiple pre-miRNA characteristics since they are not aware of the importance or existence of this relationship. However, given that pre-miRNAs can be seen as a key fitting a lock consisting of enzymes processing the sequence into a mature miRNA, the relative positions of the different characteristics are may be relevant for explaining a class prediction.

Finally, we argue that, in general, using pre-defined concepts to explain an ML system's predictions can be misleading if different users have conflicting mental models of the concepts. This mental model defines the internal representations and (semantic) meanings of the concepts based on the real-world experience of the model owner [47]. Inevitably, humans may have a different or even conflicting mental model of a particular concept. Consequently, explanations based on this concept can lead to different interpretations. Clear concept definitions and representations are required to mitigate this issue.

7.2 Results of Concept Whitening Model

Here we discuss the results obtained and derived from the concept whitening models trained with our pre-miRNA-related concepts. These results include the models' pre-miRNA classification

performance, the purity and quality of the learned concept representations, and the importance of the concepts for the classification task.

Pre-miRNA Classification Performance From the pre-miRNA classification results in section 6.2.1, we can conclude the following for the classification performance of models that contain a concept whitening layer. The forcing of a disentangled latent space in this layer can be detrimental to the performance obtained on classification tasks requiring entangled input data representations for high accuracy scores. From the results in section 6.2.1, we noticed this happens when using all six pre-miRNA concepts in the CW model, likely due to the high concept similarity and dependence. Hence, the higher this similarity and dependence, the more complicated it is to disentangle the concepts' representations, and consequently, the lower the performance on the classification task. With an improved concept set in terms of independence and discriminatory power for the classification task, we expect an increase in performance since this set likely simplifies the learning of a disentanglement CW layer's latent space.

From these results, we also found that forcing the disentanglement less rigorously using lower learning rates results in higher accuracy scores. However, this defeats the purpose of CW since it needs a disentangled latent space in the CW layer for interpretability purposes. Therefore, we conclude that CW models require a set of concepts from which independent representations can be learned to obtain high classification accuracy scores.

Concept Purity The goal of CW is to align pre-defined concepts with nodes in the CW layer (or axes of its high-dimensional latent space) such that we can use the nodes (or axes) for interpretability purposes. This requires concept representations learned in the layer to be independent, which the method aims to achieve by applying a whitening transformation to the layer's latent space. If the representations are independent, they are considered *pure* and there is no danger of *information leakage* [54]. Without purity, explanations on the influence of a *single* concept on the predictions can be misleading since they can be correlated with others. From the concept purity results given in section 6.2.1, we found that only for the *Large asymmetric bulge* (concept 4) and *At least* 90% *base pairs and wobbles in stem* (concept 2) concepts close to pure representations can be learned. For the remaining four pre-miRNA concepts (section 5.3), the results illustrated their representations are too similar. The concept example images of these concepts being too similar may have caused this issue. We were not able to collect more example images given the limited size of the *modhsa*-dataset currently used for deriving example images. However, other datasets consisting of encoded sequences complementary to those in the *modhsa*-dataset may support in finding more similar concept example images.

Another factor potentially complicating the disentanglement of concept representations in the CW layer is the whitening transformation. As explained in section 4.1, this transformation enforces disentanglement through standardization and decorrelation of the layer's latent space. However, Mahinpei et al. [54] found that decorrelating representations may not remove all statistical dependence between them. Also, Margeloiu et al. [55] found that *concept learning models* (CLMs) simultaneously learning a classification task and pre-defined concepts are encouraged to have an entangled latent space. Consequently, CLMs using this type of training are likely to suffer from impure concept representations. Moreover, Mahinpei et al. [54] argues this may always be an issue with CLMs, no matter the type of training.

Concept Learning In section 6.2.2, we discussed the model's concept learning results by visualizing the most activated images for the nodes aligned with the *At least* 90% *base pairs and wobbles in stem* and *Large asymmetric bulge* concepts, concepts 2 and 4, respectively. We also included the least and some moderately activated images for a completer view on the concept learning. In the most activated images, we highlighted image regions based on the empirical receptive field of the concepts to derive more concrete assumptions about the learning. We found it challenging to perform this derivation task if these regions are different for the same concept.

Besides the purity (or independence) issues, there are multiple reasons why CW models can have difficulties with learning pre-defined concepts. One of them is the location of the CW layer in the convolutional neural network. CNNs learn low-level patterns in the earlier layers and high-level objects in the final ones [64]. We chose to convert one of our CNN's final layers into a CW layer since our pre-miRNA concepts represent high-level objects. We could have obtained better learning results if we converted one of the earlier layers into a CW one. As Rudin et al. [64] explains, the hierarchy of concepts in the concept set should be taken into account when choosing which layer(s) to convert to CW layers. This concept hierarchy does not only refer to the distinction between low- and high-level data patterns; it can also refer to one high-level object being part of another. The same reasoning applies to low-level patterns. Although there have been some efforts in creating hierarchical concept-based explanations [40, 46], best practices are yet to be determined.

The definitions and exemplary images of the concepts can be another reason for a CW model's limited concept learning. Defining the influence of these images on the concept learning results is challenging. It may be the concept is irrelevant for the associated classification task. Or the CW model cannot identify it due to unclear example images. Taking out the unclarities by using more (similar) example images for the concept would promote evaluating which of the two cases holds.

Concept Importance We defined the importance of our pre-miRNA concepts for the class predictions using the *concept influence*-metric (5.1). This metric uses the model's computations that transform (concept) activation values from the CW layer into class predictions by weighting them in the model's output nodes linked to the two pre-miRNA classes. Consequently, we obtain two output values for an input instance, each representing the chance the instance belongs to the class linked to the value. Although the metric provides a quantitative explanation of the model's prediction for the given instance, the understandability would increase if we obtained only one output value. In general, reasoning about only one value instead of two values simultaneously is considered easier. Besides, this would decrease the amount of information included in the explanation, enabling one to focus on the most important aspects. The domain expert evaluation confirms this finding (section 6.3). We conducted several experiments with a CW model containing only one output node. However, as shown by the results given in appendix C.3, obtaining accurate classification results with this model configuration is challenging. It may be that forcing the model to generate class predictions from the output of the CW layer, which represents the whitened data representations summarized into scalars, using only one output node is too rigid. More specifically, this one output node seems to reduce the model's freedom to derive class predictions from the data representations compared to using two output nodes.

Besides our *concept influence*-metric (5.1), several alternatives exist for defining the importance of concepts for the classification task. For instance, we could use the permutation-based metric created by Chen et al. [13] which computes the ratio between the *switched loss* and the actual loss of the model. The former is calculated by randomly permuting samples along the CW layer's latent space axis aligned with a concept. However, the dependence on the sample switching makes the metric unstable with different permutations and thus less desirable than the *concept influence*-metric [59]. Similar to post-hoc concept-based interpretability method TCAV [42], we could also use gradient-based methods that measure the model's sensitivity to concept c_j based on the derivative in the direction of latent concept representation $\mathbf{Z}_{c_j}$. In Figure 4.1, the arrows pointing towards representations $\mathbf{Z}_{c_j}$ illustrate these directional derivatives. However, given we base the concept purity and learning explanations on activation values from the CW layer, it seems less optimal for understandability to use gradient-based methods instead of activation-based ones such as our *concept influence*-metric.

On a final note, we could use different representation types to explain the importance of concepts for the pre-miRNA classification task. Instead of using a bar chart showing the concepts' influences for the pre-miRNAs of both classes (section 6.2.2), we could use a decision tree to provide non-linear explanations based on the hierarchy of the importance of the concepts. In appendix C an example of such an explanation is given. Currently, the tree is not very informative since it includes only two concepts. However, we assume that adding more will increase its usefulness.

Definition of New Pre-miRNA Concepts Besides evaluating the importance of pre-defined concepts for the pre-miRNA detection task, the CW model also enables deriving new concepts from residual image information it finds important for the task. Namely, the model aligns pre-defined concepts with the axes of the CW layer's latent space. The number of space axes is equal to the number of layer nodes. If this number exceeds the number of concepts, the remaining nodes (or axes) handle residual information present in the input images. In section 6.2.2, we explained that new, potentially relevant concepts for the pre-miRNA detection task can be derived from the highly influential nodes, where the influence is defined using the *concept influence*-metric (5.1). Similar to our concept learning analysis, we visualized the images with the highest activation values for these nodes to derive which information they learned. Deriving concrete concepts definitions from these images is difficult if the shared information in them is not straightforward. Discussions with domain experts on possible concept definitions may solve this issue, emphasizing domain experts' involvement and expertise is necessary for interpretability of the specific ML application [64].

7.3 Evaluation with Domain Expert

Currently, we evaluate the framework and its results with a domain expert by interviewing the expert and focusing on the *understandability* of the explanations generated by our framework. This evaluation has provided some interesting insights into the practical utility of the framework. A more detailed evaluation could be performed using an *application-grounded* evaluation [19]. In this type of evaluation, the focus lies on evaluating the explanation quality given the intended end task of the application. In our case, an end task could be to let multiple domain experts test the framework's ability to support them in interpreting the pre-miRNA detection task results and finding (structural) pre-miRNA characteristics.

Overall, we can conclude from the domain expert evaluation that our framework is a first step towards concept-based interpretability of the pre-miRNA detection results.

7.4 Conclusions

In this chapter, we discussed several different aspects of our concept-based interpretability framework for image-based pre-miRNA detection. We found defining concepts for concept-based interpretability models before training and evaluating them on a given learning task can be challenging. More specifically, it is generally unknown which concepts ML models consider relevant for interpretability purposes. Furthermore, we concluded that using human-annotated concepts may introduce human bias, limiting the explanatory power of the interpretability model since the human perception of concepts relevant for explanation purposes may differ from that of the model.

Next, we discussed the results provided and derived from our concept whitening models. We found the model requires independence of the learned representations of the pre-defined concepts to obtain accurate classification performance. Dependence between representations can originate from many reasons, including too similar or overlapping example images for different concepts. Next, we concluded there exist many causes why a CW model may not learn a pre-defined concept. Besides independence issues, the location of the CW layer in the CNN and the hierarchy between concepts can play a role. For the concept importance explanations, we decided that activation-based methods (such as our *concept influence*-metric) are more suitable than permutation-based or gradient-based ones, given the concept purity and learning explanations are also activation-based. Finally, we discussed the derivation of new concepts from a CW model's handling of residual image information. Although possible, this can be challenging if the images and their highlighted regions from which concepts should be derived are diverging.

We ended with a brief discussion on the conducted evaluation with a domain expert. The current interview-based evaluation approach has provided some interesting insights. More extensive evaluations such as application-based ones are required to investigate the framework's usefulness in more depth.

Chapter 8

Conclusions

8.1 Concluding Summary

In this work, we focused on providing domain experts with an understanding of predictions generated by a machine learning model trained to solve the image-based precursor microRNA detection task. We set out to answer the following research question: *How can we provide domain experts with global and local concept-based explanations of (potentially) discriminative characteristics present in the positive and negative pre-miRNA sequences encoded as images?*

We initiated the design of our framework by defining our motivation for interpretability based on a literature review. Next, we chose an interpretability method most appropriate given this motivation and the context of the image-based pre-miRNA detection task, also based on a literature review. Based on the method's requirements, we extended our framework with a concept definition step preceding the step of applying the method. The final steps of the framework include an evaluation of the obtained classification and interpretability results (together with a domain expert).

8.1.1 Literature Review

Motivation for Interpretability We initiated the design of our framework by defining our motivation for interpretability to answer our first subquestion: *What is the main motivation for requiring interpretability of our image-based pre-miRNA detection application?*

Based on a literature review, we found the *promotion of knowledge discovery* motivation most suitable for our goal of explaining the pre-miRNA class prediction based on their structural characteristics.

Interpretability Method Next, we extended our literature review by analyzing state-of-the-art interpretability methods matching our motivation and problem context. More specifically, we addressed our second subquestion: *Given the context of the image-based pre-miRNA detection problem and the interpretability required to promote discovery of pre-miRNA-related knowledge, which interpretability method is most appropriate?*

First, we compared several existing approaches for providing ML interpretability. We've established that concept-based interpretability can be a suitable approach for the task at hand. Contrary to (groups of) pixels commonly used for explaining predictions, concepts have a semantic meaning, making them meaningful units for generating *human-understandable* explanations. Moreover, the discriminative pre-miRNA features we intended to use for interpreting the image-based pre-miRNA classification results may be seen as concepts.

Focusing on concept-based interpretability, we found two main approaches in the literature, one requiring pre-defined concepts and the other deriving them from input data automatically. We found the first approach more convenient since the concepts derived in the latter can be difficult to interpret, conflicting with our goal of providing *understandable* explanations. Additionally, the

fact that concept-based explanations given by post-hoc methods can be misleading while intrinsic methods are *faithful by design* made us conclude intrinsic methods are more appropriate. Among the reviewed intrinsic methods, we found *Concept Whitening* (CW) [13] best fitting with our interpretability motivation and goal of providing *understandable* concept-based explanations.

8.1.2 Concept Whitening for Image-based Pre-miRNA Detection

Requirements Next, we explained the chosen concept-based interpretability method in more detail. The method requires an existing convolutional neural network (CNN) and a set of pre-defined concepts complementary to the dataset used for learning the classification task. We chose the CNN trained by Cordero et al. [16] on the encoded pre-miRNA instances defining the classification dataset of interest. Since this dataset does not include concepts, the next step was to define concepts in the form of example images. Leveraging existing domain knowledge on (pre-)miRNA detection, we formulated concept definitions and linked them to the encoded data sequences. Finally, we refined the definitions of our six main concepts based on their presence and correlation with others in the classification dataset.

Provided Interpretability Since CW models learn concepts from example images, the *purity* and *quality* of their learned representations should be analyzed before providing concept-based explanations for the class predictions. Purity is required to generate faithful explanations, while the quality informs model users whether the representations are similar to their mental model of the concepts. The concept-based prediction explanations are generated based on the activation values of model parts trained to recognize the concepts in the input images. Since we applied some adjustments to the CW method, we simplified the generation of concept-based explanations. Namely, we facilitated directly quantifying the importance of concepts learned concepts for detecting pre-miRNAs in the images.

8.1.3 Evaluation

Results CW models simultaneously learn a classification task and pre-defined concepts to use the latter to explain the classification predictions. Therefore, we presented results for both learning tasks. We concluded (close to) pure concept representations are required to obtain accurate classification performance scores. Consequently, we found only two of our six pre-defined concepts usable for the method.

The model's learned concept representations illustrated that one concept is learned in line with expectations. Also, we found this concept to positively influence the classification of *positive* pre-miRNAs, and vice versa for *negative* pre-miRNAs, confirming previous pre-miRNA detection findings. The other concept seems unimportant for explaining the pre-miRNA classification results. Finally, we found the model considers a pattern related to another previous pre-miRNA detection finding important for the positively classified sequences from its residual image information handling.

Perceived Usefulness of Framework Together with a domain expert, we evaluated the framework and its results to answer our final subquestion: *How useful are local and global concept-based explanations provided by an interpretability method for supporting domain experts in finding discriminative pre-miRNA characteristics?*

We interviewed the domain expert using questions focusing on the usefulness of the proposed framework in general and the *understandability* of the concept-based explanations. The collected feedback showed the framework has the potential to become useful in practice. Its feature allowing the confirmation of previous pre-miRNA detection findings and the derivation of new pre-miRNA-related concepts from residual information handling is considered valuable. However, the flawed understandability of the explanations has a detrimental effect on the framework's usefulness. More specifically, their distinction between the two pre-miRNA classes is challenging to comprehend. Since this distinction originates from the architecture of the existing CNN used in our framework, future work should prioritize creating a CW model without this distinction.

8.2 Limitations and Future Work

Finally, we discuss some limitations of our framework and possible solutions. Firstly, the dependence on pre-defined concepts is both an advantage and a disadvantage. Concepts created by humans have a semantic meaning, making them meaningful units for creating *human-understandable* explanations for predictions. On the other hand, human concept annotators are likely to introduce human bias into the explanation process. Their perception of concepts useful for interpretability purposes may be different than that of an ML model. Moreover, we realized that defining concepts relevant to explain predictions before training and evaluating the classification model generating the predictions is challenging. In general, it seems counter-intuitive to force deep learning models, known to derive features from the input data to solve a given task, to learn features (or concepts) humans consider relevant for explaining the model's predictions.

Second, we noticed our results are highly dependent on the purity (or independence) of the learned concept representations. Our current concept set seems to contain several concepts whose representations are too similar. These issues may stem from various reasons, such as unclear concept definitions, overlapping example images for different concepts or the presence of a certain hierarchy in the concepts. Although efforts to solve these issues may be worthwhile, concept independence issues may always be present in models learning pre-defined concepts. However, since concept-based explanations can be misleading if different users have conflicting mental models of these concepts, ambiguities in concept definitions or representations should be avoided at any time.

Given these issues with pre-defined concepts, one could choose to use the generative disentanglement learning methods explained in section 3.3.2. These methods derive latent factors from the input data, which can be regarded as concepts. However, interpreting the learned latent factors can be challenging and sometimes even impossible. We can make similar conclusions for the feature of the CW method that enables deriving new concepts from the residual image information handling in the CW layer. However, we believe this feature and the generative methods can provide relevant concept-related information for the pre-miRNA detection problem. Therefore, we suggest future work to focus on using the disentanglement learning methods or further analyzing the residual information learned in the CW layer to define new concepts. Using these new concepts in a CW model can support quantifying their relevance for the pre-miRNA detection task can be quantified and observing further refinements for their definitions. We imagine this iterative or *active learning* process, where new concepts are found from residual pre-miRNA image information and existing ones are extensively validated, to be extremely valuable.

Bibliography

[1] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

[2] Jens Allmer. "Computational and bioinformatics methods for microRNA gene prediction". In: *MiRNomics: MicroRNA biology and computational analysis*. Springer, 2014, pp. 157–175.

[3] Francesco Angelucci et al. "MicroRNAs in Alzheimer's disease: diagnostic markers or therapeutic agents?" In: *Frontiers in pharmacology* 10 (2019), p. 665.

[4] André Araujo, Wade Norris and Jack Sim. "Computing receptive fields of convolutional neural networks". In: *Distill* 4.11 (2019), e21.

[5] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[6] Youhuang Bai et al. "Toward a next-generation atlas of RNA secondary structure". In: *Briefings in bioinformatics* 17.1 (2016), pp. 63–77.

[7] Yoshua Bengio, Aaron Courville and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[8] M Bhaskaran and M Mohan. "MicroRNAs: history, biogenesis, and their evolving role in animal development and disease". In: *Veterinary pathology* 51.4 (2014), pp. 759–774.

[9] Nadia Burkart and Marco F Huber. "A survey on the explainability of supervised machine learning". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[10] Rich Caruana et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 1721–1730.

[11] Diogo V Carvalho, Eduardo M Pereira and Jaime S Cardoso. "Machine learning interpretability: A survey on methods and metrics". In: *Electronics* 8.8 (2019), p. 832.

[12] Chaofan Chen et al. "This looks like that: deep learning for interpretable image recognition". In: *CoRR* abs/1806.10574 (2018). arXiv: 1806.10574. URL: http://arxiv.org/abs/1806.10574.

[13] Zhi Chen, Yijie Bei and Cynthia Rudin. "Concept whitening for interpretable image recognition". In: *Nature Machine Intelligence* 2.12 (2020), pp. 772–782.

[14] Michael Chromik and Martin Schuessler. "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI." In: *ExSS-ATEC@ IUI*. 2020.

[15] European Commission. *General Data Protection Regulation*. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[16] Jorge Cordero, Vlado Menkovski and Jens Allmer. "Detection of pre-microRNA with Convolutional Neural Networks". In: *bioRxiv* (2020), p. 840579.

[17] Enrico De Smaele, Elisabetta Ferretti and Alberto Gulino. "MicroRNAs as biomarkers for CNS cancer and other disorders". In: *Brain research* 1338 (2010), pp. 100–111.

[18] Binh Thanh Do et al. "Precursor microRNA identification using deep convolutional neural networks". In: *BioRxiv* (2018), p. 414656.

[19] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[20] AE Erson and EM Petty. "MicroRNAs in development and disease". In: *Clinical genetics* 74.4 (2008), pp. 296–306.

[21] Feng-Lei Fan et al. "On interpretability of artificial neural networks: A survey". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* (2021).

[22] Wenwen Fang and David P Bartel. "The menu of features that define primary microRNAs and enable de novo design of microRNA genes". In: *Molecular cell* 60.1 (2015), pp. 131–145.

[23] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[24] Amirata Ghorbani et al. "Towards automatic concept-based explanations". In: *arXiv preprint arXiv:1902.03129* (2019).

[25] Leilani H Gilpin et al. "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[26] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI magazine* 38.3 (2017), pp. 50–57.

[27] Yash Goyal et al. "Explaining classifiers with causal concept effect (cace)". In: *arXiv preprint arXiv:1907.07165* (2019).

[28] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[29] Jinju Han et al. "The Drosha-DGCR8 complex in primary microRNA processing". In: *Genes & development* 18.24 (2004), pp. 3016–3027.

[30] David A Hendrix. "Applied Bioinformatics". In: (2019).

[31] Bernease Herman. "The promise and peril of human evaluation for model interpretability". In: *arXiv preprint arXiv:1711.07414* (2017), p. 8.

[32] Ivo L Hofacker. "Vienna RNA secondary structure server". In: *Nucleic acids research* 31.13 (2003), pp. 3429–3431.

[33] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[34] Xinhua Ji. "The mechanism of RNase III action: how dicer dices". In: *RNA interference* (2008), pp. 99–116.

[35] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[36] Wenjing Kang and Marc R Friedländer. "Computational prediction of miRNA genes from small RNA sequencing data". In: *Frontiers in bioengineering and biotechnology* 3 (2015), p. 7.

[37] Dmitry Kazhdan et al. "Is Disentanglement all you need? Comparing Concept-based & Disentanglement Approaches". In: *arXiv preprint arXiv:2104.06917* (2021).

[38] Dmitry Kazhdan et al. "Now You See Me (CME): Concept-based Model Extraction". In: *arXiv preprint arXiv:2010.13233* (2020).

[39] Agnan Kessy, Alex Lewin and Korbinian Strimmer. "Optimal whitening and decorrelation". In: *The American Statistician* 72.4 (2018), pp. 309–314.

[40] Mohammed Khaleel et al. "Hierarchical Visual Concept Interpretation for Medical Image Classification". In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2021, pp. 25–30.

[41] Been Kim et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)". In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[42] Tae Wan Kim. "Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test". In: *arXiv preprint arXiv:1810.09598* (2018).

[43] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[44] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1885–1894.

[45] Pang Wei Koh et al. "Concept bottleneck models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5338–5348.

[46] Avinash Kori et al. "Abstracting deep neural networks into concept graphs for concept level interpretability". In: *arXiv preprint arXiv:2008.06457* (2020).

[47] Todd Kulesza et al. "Too much, too little, or just right? Ways explanations impact end users' mental models". In: *2013 IEEE Symposium on visual languages and human centric computing*. IEEE. 2013, pp. 3–10.

[48] Isaac Lage et al. "An evaluation of the human-interpretability of explanation". In: *arXiv preprint arXiv:1902.00006* (2019).

[49] Shaohua Li et al. "Mismatched and wobble base pairs govern primary microRNA processing by human Microprocessor". In: *Nature communications* 11.1 (2020), pp. 1–17.

[50] Yitong Li et al. "Targeting EEG/LFP synchrony with neural nets". In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 4620–4630.

[51] Zachary C Lipton. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[52] Francesco Locatello et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[53] Francesco Locatello et al. "Weakly-supervised disentanglement without compromises". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6348–6359.

[54] Anita Mahinpei et al. "Promises and Pitfalls of Black-Box Concept Learning Models". In: *arXiv preprint arXiv:2106.13314* (2021).

[55] Andrei Margeloiu et al. "Do Concept Bottleneck Models Learn as Intended?" In: *arXiv preprint arXiv:2105.04289* (2021).

[56] Damien de Mijolla et al. "Human-interpretable model explainability on high-dimensional data". In: *arXiv preprint arXiv:2010.07384* (2020).

[57] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[58] Yao Ming et al. "Interpretable and steerable sequence learning via prototypes". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 903–913.

[59] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. 2019.

[60] Tuan Anh Nguyen et al. "Functional anatomy of the human microprocessor". In: *Cell* 161.6 (2015), pp. 1374–1387.

[61] Matthew O'Shaughnessy et al. "Generative causal explanations of black-box classifiers". In: *arXiv preprint arXiv:2006.13913* (2020).

[62] Ribana Roscher et al. "Explainable machine learning for scientific insights and discoveries". In: *IEEE Access* 8 (2020), pp. 42200–42216.

[63] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[64] Cynthia Rudin et al. "Interpretable machine learning: Fundamental principles and 10 grand challenges". In: *arXiv preprint arXiv:2103.11251* (2021).

[65] Stefan Rüping et al. "Learning interpretable models". In: (2006).

[66] Müşerref Duygu Saçar and Jens Allmer. "Machine learning methods for microRNA gene prediction". In: *miRNomics: MicroRNA Biology and Computational Analysis*. Springer, 2014, pp. 177–187.

[67] Amitojdeep Singh, Sourya Sengupta and Vasudevan Lakshminarayanan. "Explainable deep learning models in medical image analysis". In: *arXiv preprint arXiv:2005.13799* (2020).

[68] Karen Smith. *Orthogonal Transformations*. 2015.

[69] Kacper Sokol and Peter Flach. "Explainability fact sheets: a framework for systematic assessment of explainable approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 56–67.

[70] Ning Xie et al. "Explainable deep learning: A field guide for the uninitiated". In: *arXiv preprint arXiv:2004.14545* (2020).

[71] Bing Xu et al. "Empirical evaluation of rectified activations in convolutional network". In: *arXiv preprint arXiv:1505.00853* (2015).

[72] Fan Yang, Mengnan Du and Xia Hu. "Evaluating explanation without ground truth in interpretable machine learning". In: *arXiv preprint arXiv:1907.06831* (2019).

[73] Chih-Kuan Yeh et al. "On concept-based explanations in deep neural networks". In: (2019).

[74] Mohammad Nokhbeh Zaeem and Majid Komeili. "Cause and Effect: Concept-based Explanation of Neural Networks". In: *arXiv preprint arXiv:2105.07033* (2021).

[75] Bolei Zhou et al. "Interpreting deep visual representations via network dissection". In: *IEEE transactions on pattern analysis and machine intelligence* 41.9 (2018), pp. 2131–2145.

Appendix A

Concept Whitening Method

In the following sections, the different techniques used in the concept whitening method [13] are explained in more detail. First, we describe whitening and orthogonal transformations in general. Next, we give more details on the explanation generation process for the pre-miRNA detection task using the *concept influence*-metric 5.1.

A.1 Whitening Transformation

A whitening procedure linearly transforms a vector of random variables with known positive definite covariance matrix into a new random vector such that the covariance matrix is equal to the identity matrix. More specifically, d-dimensional input vector $\mathbf{x} = (x_1, ..., x_d)^T$ with positive definite covariance matrix $var(\mathbf{x})$ is transformed into random vector $\mathbf{z} = (z_1, ..., z_d)^T = \mathbf{W}\mathbf{x}$ with covariance matrix $var(\mathbf{z}) = \mathbf{I}$. The procedure is referred to as whitening, since the covariance matrix of the input $\mathbf{x}$ is transformed to a *white* covariance matrix. As a result, the input vector $\mathbf{x}$ is transformed to *white noise* vector $\mathbf{z}$. After whitening, the random variables of input vector $\mathbf{x}$ are less correlated with each other, and they all have the same variance [39].

The whitening matrix $\mathbf{W}$ is a $d \times d$ matrix. For this matrix, the following constraints should hold. We have that $\mathbf{W}\mathbf{W}^T = \mathbf{I}$ since $var(\mathbf{z}) = \mathbf{I}$, from which follows $\mathbf{W}(\mathbf{W}\mathbf{W}^T) = \mathbf{W}$. This condition holds if $\mathbf{W}^T\mathbf{W} = \Sigma^{-1}$. Since it is not unique there exist many different whitening methods, such as PCA whitening and Cholesky whitening [39].

A.2 Orthogonal Transformation

Similar to a whitening transformation, an orthogonal transformation is a linear transformation. With an orthogonal one applied to a matrix consisting of a set of vectors, the goal is to preserve the length of and angles between the vectors [68]. More specifically, consider the case where we apply an orthogonal transformation $\mathbb{R}^n \xrightarrow{T} \mathbb{R}^n$ to all vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$. The result is as follows:

$$\mathbf{u} \cdot \mathbf{v} = T\mathbf{u} \cdot T\mathbf{v} \tag{A.1}$$

such that $|T\mathbf{u}| = |\mathbf{u}|$ and $|T\mathbf{v}| = |\mathbf{v}|$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. As a result, the transformation enables rotating or flipping a matrix of vectors while preserving the geometry of the matrix.

Appendix B

Concepts

In this chapter, we provide more information on the concept generation process. We introduce our preliminary concepts and explain how they are refined into the concepts given in section 5.3.1.

B.1 Preliminary Concepts

As explained in section 5.3, the pre-miRNA-related concepts are defined based on previous (pre-)miRNA detection results, including the saliency results of DeepMir [16], and the used pre-miRNA encoding algorithm. We assume this improves the domain experts' understandability of the concepts and the concept-based explanations for the pre-miRNA detection results. Since the number of validated discriminative structural characteristics of pre-miRNA is low, defining relevant concepts is challenging. Consequently, we defined many different ones and refined them based on their presence and correlation in the images of the two pre-miRNA classes in the *modhsa*-dataset.

B.1.1 Concept Definition and Annotation

Examples of our first concepts are the number or shape of (a)symmetric bulges in the non-coding RNA sequence or the pixel color most present in a bulge. These concepts are either too general or difficult to annotate in the encoded sequences. Therefore, besides the two motif concepts (concepts 5 and 6) defined in section 5.3.1, we ended up with the following four preliminary and easy to annotate concepts. We provide formal definitions, constraints for annotating them in the encoded sequences, and a link to a figure showing examples. Namely, Figure B.1(a) shows a subset of the concepts denoted in a encoded pre-miRNA sequence of the positive class, in Figure B.1(b) this is done in an instance of the negative class.

Presence of a Terminal Loop All pre-miRNA sequences should contain a terminal loop in order to have a hairpin-shape, while other non-coding RNA sequences may contain one. Hence, we assumed a concept specifying the presence of this terminal loop in the sequence should be in favor of the positive pre-miRNA class. The concept is more formally defined below.

Concept 7 ***(Presence of a terminal loop)***

- *Definition: Binary concept specifying whether the sequence contains a terminal loop.*
- *Annotation constraints: A terminal loop is recognized in the images in case the first pair of the sequence from left to right does not contain a gap. A gap is represented by two black pixels stacked on top of each other.*
- *Example: Figure B.1(a)1.*

Base Pairs and Wobbles in Pre-miRNA Stem From previous pre-miRNA detection results, we found base pairs (and wobbles) in the stem are preferred. We defined several concepts related to this characteristic, such as the number of base pairs or wobbles in a specific stem location. In the end, we decided on a more general concept, namely the concept that counts the number of base pairs and wobbles in the complete stem. We defined the concept more formally as follows.

Concept 8 ***(Frequency of base pairs and wobbles in the stem)***

- *Definition: A continuous concept with range* $[0, 1]$ *defining the fraction of base pairs and wobbles over all stem pairs.*
- *Annotation constraints: For each pair in the sequence stem, the combination of colors and the length of the bars is analyzed. In case of a blue-red or red-blue color combination with bars consisting of 2 pixels, a C-G or G-C pair is identified. A combination of yellow-green or green-yellow bars of with length 3 pixels each defines a A-U or U-A pair. Finally, wobbles (G-U or U-G pairs) are recognized in case of red-green or green-red bars consisting of 4 pixels. The total amount of recognized pairs that adhere to this constraint is divided by the total stem length to get the frequency score. The total stem length is defined as the number of pixels from the end of the terminal loop until the whitespace area in the right part of the image. The loop end is defined as the first base pair or wobble after the loop, which consists of pairs with bar lengths exceeding 3 pixels each.*
- *Example: Figure B.1(a)2.*

Presence of an Asymmetric Bulge instead of a Terminal Loop Related to concept 7, we defined a concept assumed in favor of the negative pre-miRNA class. Namely, this concept specifies whether the RNA sequence contains an asymmetric bulge on the location where a terminal loop should be. Hence, in case the sequence does not contain a terminal loop (which is by definition a symmetric bulge), it includes this concept. The formal definition of the concept is as follows.

Concept 9 ***(Presence of an asymmetric bulge instead of a terminal loop)***

- *Definition: Binary concept defining the presence of an asymmetric bulge instead of terminal loop.*
- *Annotation constraints: Given a terminal loop is identified if the leftmost sequence pair is not a gap, all sequences that do have a gap as leftmost pair contain this concept.*
- *Example: Figure B.1(b)3.*

Presence of a High Asymmetric Bulge Previous pre-miRNA detection results have shown that base pairs and wobbles are preferred. Also, the number of asymmetric bulges sized similar to the terminal loop should be minimized. Therefore, we defined a concept representing the presence of a high asymmetric bulge in the sequence with the following formal definition.

Concept 10 ***(Presence of a high asymmetric bulge)***

- *Definition: Binary concept specifying whether the presence of a high asymmetric bulge.*
- *Annotation constraints: The asymmetric bulge itself is recognized by a sequence of consecutive gaps, which are black colored pixels. The asymmetric bulge is considered high in case the highest bar contained in the bulge reaches the border of the image. Hence, in case the colored pixels of this bar continue until the border of the image is reached.*
- *Example: Figure B.1(b)4.*

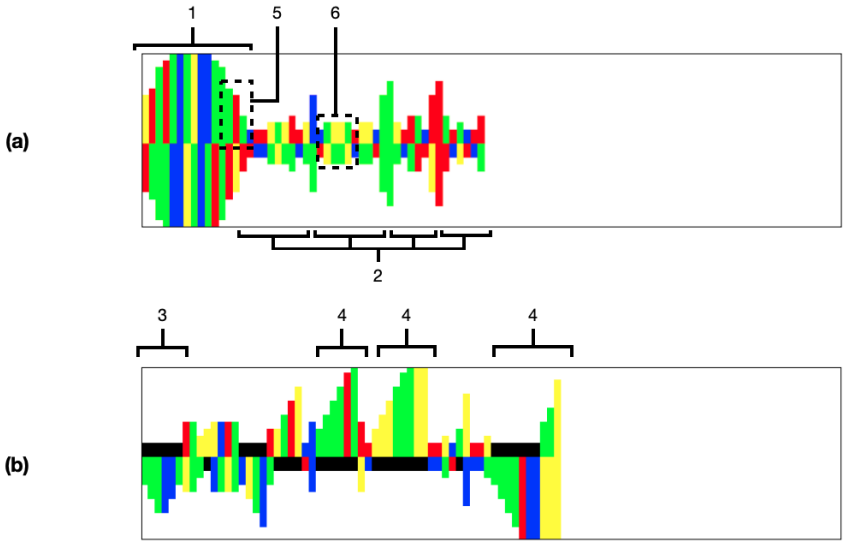

Figure B.1: Graphical representation of the six preliminary concepts in the encoded pre-miRNA sequences. In (a), the *terminal loop* (1), *frequency of base pairs and wobbles in stem* (2), *U-G-U motif* (3), and *A-U pairs motif* (4) concepts are shown in the encoded *hsa-mir-26a-2* pre-miRNA. In (b), the remaining two concepts, the *asymmetric bulge instead of a terminal loop* (3) and *high asymmetric bulge* (4) concept are shown in the *hsa1_21035* pre-miRNA.

B.1.2 Concept Quantitation

Concept Presence With these concept definitions and annotation constraints, the annotation process of the images in the *modhsa*-dataset can start. Afterwards, the presence of the concepts in the dataset can be analyzed. In Table B.1 the statistics on the presence of the binary concepts in the *modhsa*-dataset are given. The statistics show that the terminal loop and two motifs are more present in the positive class than in the negative. Almost all positively labeled images contain a terminal loop, while for the negative class this is just over half of the images. According to the domain expert, the positively classified pre-miRNA instances not containing a terminal loop are probably mistakes in the database. Since concept 9 is the opposite of concept 7, the percentages of these two add up to 100% for both classes. Finally, the large asymmetric bulge concept is recognized in almost 70% of the negatively labeled instances, and in only 17% of the positive ones.

The presence of the remaining concept of our six main concepts, the frequency of base pairs and wobbles in the stem concept, is defined by measuring the average frequency score. The last row in Table B.1 shows the average frequency and associated standard deviation for both classes. The scores show that on average the fraction of base pairs and wobbles in the stem of positive instances is 0.73, while for the negative ones this is only 0.34.

Concept	**Positive class**	**Negative class**
Presence of terminal loop	97%	53%
Average (± std) frequency of base pairs and wobbles in stem	0.73 (±0.14)	0.34 (±0.19)
Presence of asymmetric bulge instead of terminal loop	3%	47%
Presence of high asymmetric bulge	17%	69%
Presence of U-G-U motif	26%	9%
Presence of A-U pairs motif	40%	11%

Table B.1: Presence of the five binary concepts and average and standard deviation of the non-binary concept in the positive and negative class of the *modhsa*-dataset.

Concept Correlation Besides the presence of the concepts in the *modhsa*-dataset, we also analyzed the correlation between all concepts. In Figure B.2 the resulting correlation values for the six concepts are given. We find the *Terminal loop* and *Frequency base pairs and wobbles in stem* concepts highly correlated, with a value of 0.81. We find these two concepts to also have moderate correlation values with other concepts, ranging between 0.27 and 0.44. Several concepts have negative correlation values with other concepts. In general, these values occur between concepts related to *positive* pre-miRNAs and concepts related to *negative* pre-miRNAs.

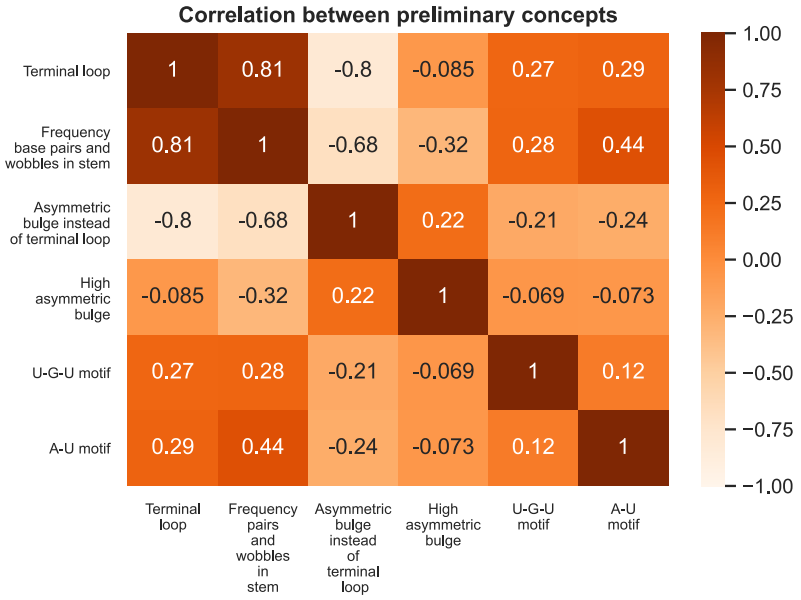

Figure B.2: Correlations between the six preliminary pre-miRNA concepts. The darker the orange color, the more positive the correlation. The lighter, the more negative.

B.2 Concept Refinement

For the concepts that are quite general, being concepts 7, 8, 9, and 10, we refined their definitions by on adding some specifications to them. These specifications are based on factors linked to the appearance of the concepts such as the width or height of bulges. We collected statistics such as the averages and standard deviations of these factors for both classes and used these values to add further constraints to the concepts.

Appendix C

Evaluation

In this chapter, we provide more results provided and derived by our concept whitening models. Also, the questions used in the interview with the domain expert to evaluate the framework and results are given.

C.1 Concept learning

In section 6.2.2, we illustrated the learning of the pre-defined concepts done by the CW model using the most, least and some moderately activated images. Also, the images included the normalized empirical receptive fields of the concepts of interest. In this section, we include other highly, moderately and low activated images with these receptive fields.

- Concept 2: *At least* 90% *base pairs and wobbles in stem*

  In section 6.2.2 Figure 6.11 the top three most activated test set images for the node in the CW layer aligned with the concept are given. In Figure C.1 the other seven images completing the top 10 most activated ones are shown. The images include the receptive field of concept 2. These images also show the model has learned a representation for the concept in line with our expectations. The receptive fields illustrate the model considers the stem location where most base pairs are located *highly* similar to the learned concept representation. Interestingly, wobbles are only included in the first and third images in the first row, indicating the model does not find them very representative of the concept.

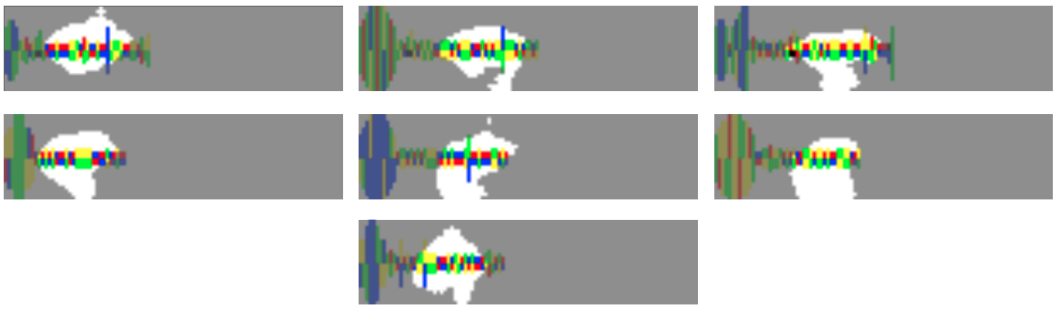

Figure C.1: Images on place four to ten in the top ten most activated images for concept *At least* 90% *base pairs and wobbles in stem* (concept 2) including its receptive field.

Next, we provide more examples of the moderately and least activated test set images for concept 2. In Figure C.2, we provide more examples of moderately activated test set images. We find most contain several (a)symmetric bulges in the stem instead of base pairs or wobbles. In section 6.2.2 Figure 6.14 we showed the top three least activated test set images for concept 2. Therefore, in Figure C.3 the remaining images of the top 10 least activated ones are given. Again, these images mostly contain (a)symmetric bulges in the stem area. Interestingly, comparing these images to the moderately activated ones in Figure C.2, we find that some of the least activated images contain more base pairs (and wobbles) in the stem. However, both the moderately and least activated images show the model considers (a)symmetric bulges in the stem not representative for concept 2, which is what we expected.

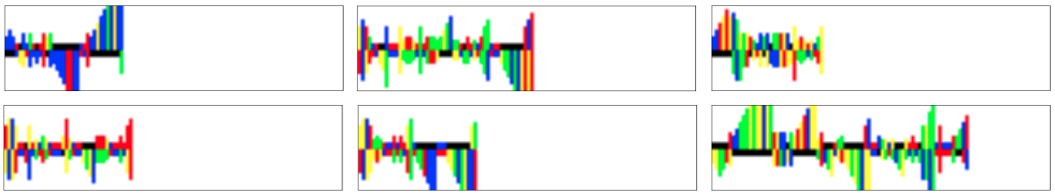

Figure C.2: Moderately activated images for concept *At least* 90% *base pairs and wobbles in stem* (concept 2).

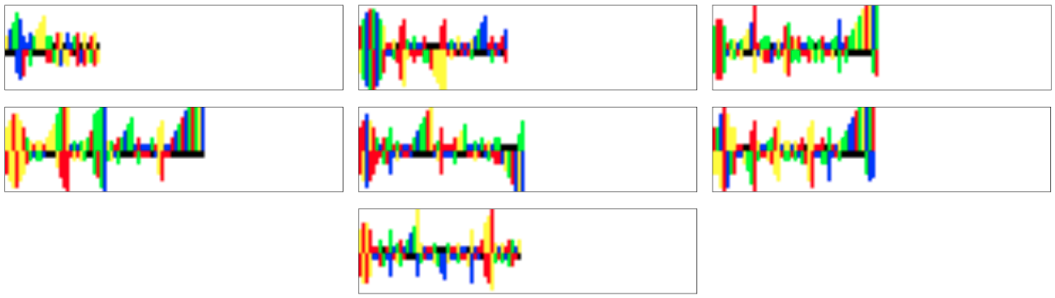

Figure C.3: Images on place four to ten in the top ten least activated images for concept *At least* 90% *base pairs and wobbles in stem* (concept 2).

- Concept 4: *Large asymmetric bulge*

  Again, we start with the other most activated images for the node in the CW layer aligned with concept 4. In Figure C.4 we provide the images on place four up until ten in the top ten most activated test set images. We find all images to contain at least one large asymmetric bulge. Similar to the findings in section 6.2.2, we find that the node aligned with the concept seems to focus on the lower right corner of the sequences. In most images, the focus is on the lower image part. In some of them, the focus seems very little and on the whitespace. Since we normalized the receptive fields representing this focus, this means the activations for this particular image for the node aligned with concept 2 may have been higher, resulting in rather weird receptive field values for concept 4. Not normalizing the values would probably result in the focus being also on the sequence and not just the whitespace. Overall, we may conclude the most activated images indeed contain our mental model of concept 4. However, since the receptive fields can differ quite a lot, especially compared to those for concept 2, it is difficult to derive concrete meanings on the focus of the node aligned with concept 4.

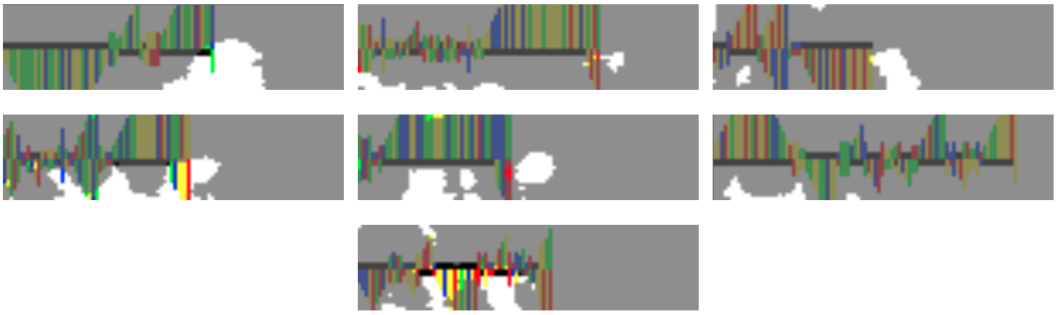

Figure C.4: Images on place four to ten in the top ten most activated images for concept *Large asymmetric bulge* (concept 4) including its receptive field.

We end this section with the moderately and least activated images for this concept. In Figure C.5, we provide some of the moderately activated test set images for the node in the CW layer aligned with concept 4. We see some images contain asymmetric bulges, but overall these can be considered small, especially when compared with those in Figure C.4. In Figure C.6, the images that complete the top ten least activated test set images besides those in Figure 6.18. These images also contain some asymmetric bulges, but again these are quite small. Comparing them to the moderately activated ones (Figure C.5, we see some of the least activated images contain more base pairs and wobbles in the stem area. This is in line with expectations, as these types of pairs can never be part of a bulge.

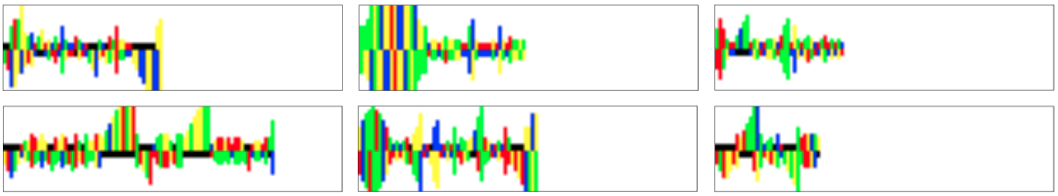

Figure C.5: Moderately activated images for concept *Large asymmetric bulge* (concept 4).

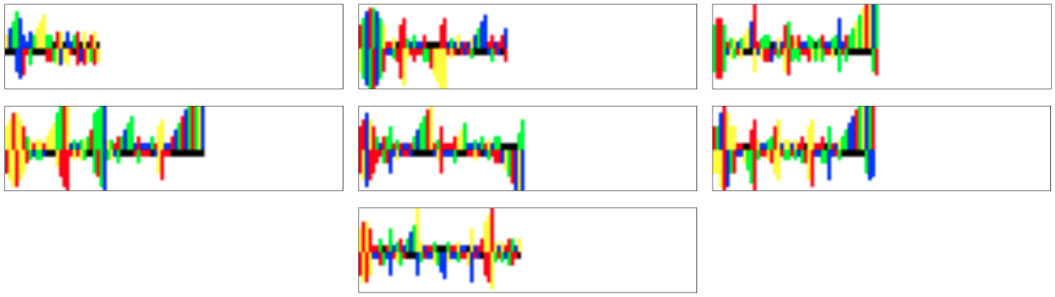

Figure C.6: Images on place four to ten in the top ten least activated images for concept *Large asymmetric bulge* (concept 4).

C.2 Concept importance

Local Explanations As explained in section 6.2.2, more local concept importance explanations are included in this section. First, we give more examples of correctly classified instances. Afterwards, we show some examples of instances that are not correctly classified by our CW model.

- Consider the input image, evidence for the pre-miRNA classes derived from the model and the final class prediction generated from the evidence given in Figure C.7. We see the instance is classified as *positive* pre-miRNA due to the similarity with the learned representation of concept 2 and the information learned in the remaining nodes of the CW layer. The similarity with concept 2 decreases the chance of being a *negative* pre-miRNA and increases the evidence for the *positive* pre-miRNA class. We find that concept 4 has little to no influence on the predictions.
- In Figure C.8 we provide an input image, evidence for the two pre-miRNA classes derived from the model, and the interpretation of this evidence leading to the final class prediction. The input image contains several asymmetric bulges, of which one reaches the image border. The presence of this bulge does not lead to a notable influence of concept 4 for both pre-miRNA classes. The similarity of the input image with the learned representation of concept 2 decreases the chance of being a *negative* pre-miRNA. However, the remaining nodes in the CW layer have learned information in favor of the negative class, increasing the evidence for the *negative* class and decreasing that of the *positive* one. As a result, the final class prediction is in favor of the *negative* class.

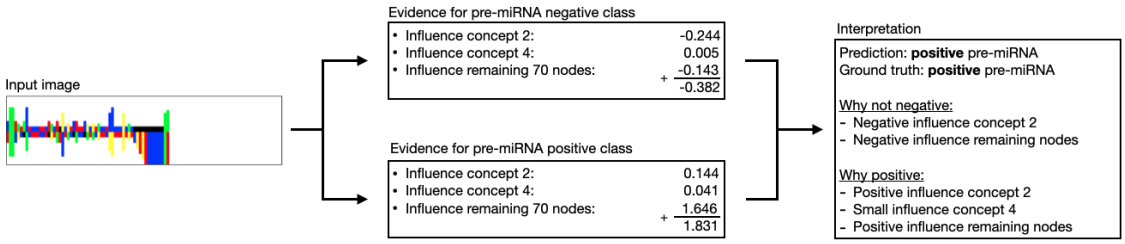

Figure C.7: Local concept influence explanation for sequence *hsa-mir-602* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

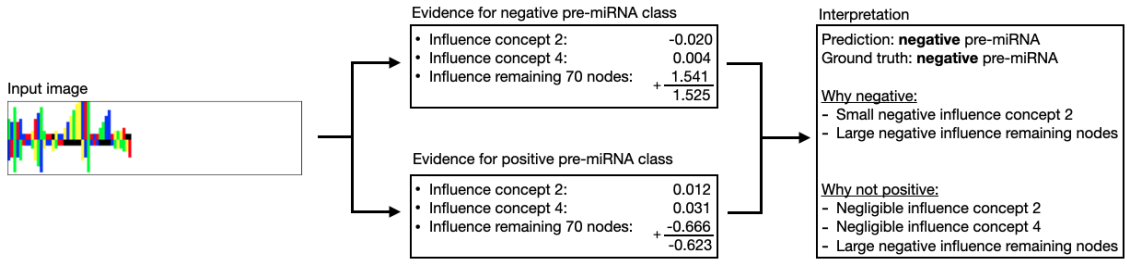

Figure C.8: Local concept influence explanation for sequence *1011* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

- Consider the input image, evidence for the pre-miRNA classes derived from the model, and the final class prediction generated from the evidence given in Figure C.9. The input image contains a large asymmetric bulge, some symmetric bulges, and some base pairs and wobbles in the stem. This latter characteristic has caused the evidence for the *negative* pre-miRNA class to decrease. However, we find the remaining nodes in the CW layer to have a large positive influence on the evidence of this class and a large negative influence on the evidence for the *positive* class. As a result, the final class prediction is in favor of the *negative* class. Interestingly, this prediction is wrong, as the image should be classified as *positive* pre-miRNA. Hence, the remaining nodes of the CW layer have learned patterns that should definitely be analyzed to evaluate the reason for this wrong prediction.
- The final local explanation given in this section is shown in Figure C.10. Again, this figure includes the input image, class evidence values which are interpreted to explain the final class prediction. We see the input image contains several base pairs and some wobbles in its stem, and some small asymmetric bulges. The base pairs and wobbles in the stem are recognized by the node aligned with concept 2, decreasing the chance of being a *negative* pre-miRNA. This similarity has increases the evidence for the *positive* pre-miRNA class. The remaining nodes in the CW layer have a large negative influence on the *negative* class and a large positive influence on the *positive* one. As a result, the final class prediction is in favor of the *positive* class. This prediction is wrong, as the ground truth label associated with the input image is *negative* pre-miRNA. Again, we realize the information learned in the remaining nodes of the CW layer, which has a large influence on the predictions, should be analyzed to understand why this prediction is wrong.

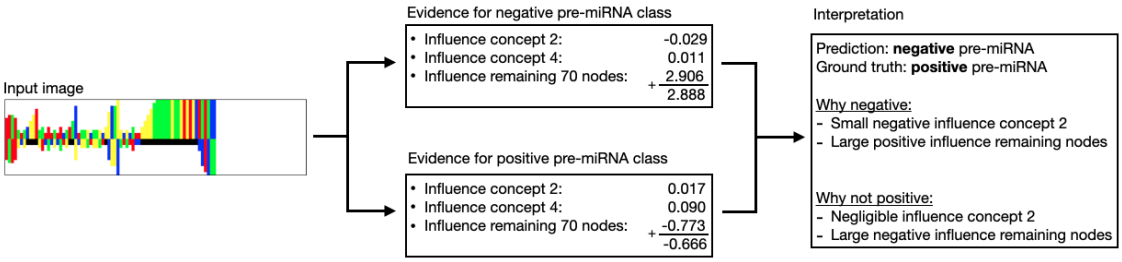

Figure C.9: Local concept influence explanation for sequence *1048* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

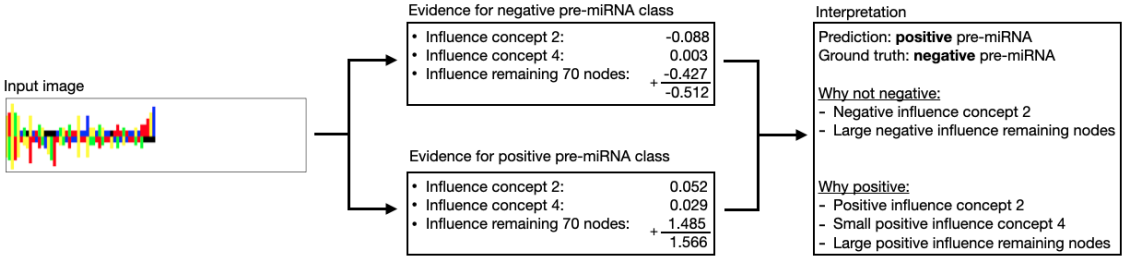

Figure C.10: Local concept influence explanation for sequence *1048* using the *At least* 90% *base pairs and wobbles in stem* (concept 2) and *Large asymmetric bulge* (concept 4) concepts.

Global Explanations In section 6.2.2, we presented global explanations for the importance of concepts for the pre-miRNA predictions using bar charts showing the average influence of the concepts on predictions made using the test set images. We can also put more focus on the activation values of the concepts that are weighted with the class weights to calculate the concept influence values. Consider Figure C.11 showing all influence values (or weighted activation values) for the two concepts and the remaining 70 nodes obtained for the positively classified instances of the *modhsa* test set. The values are colored based on their original activation value. High activations are purple and low ones are orange. The values show that concept 4 has very little influence on the predictions of the positive class. Concept 2 is more influential. Also, the higher the original activation value, the higher the influence. Finally, the sum of the concept influence values of the remaining 70 nodes is much larger than the influences of the two concepts. For the negatively classified test instances, this is similar, as illustrated in Figure C.12. Interestingly, the influence of concept 2 is slightly negative for some instances.

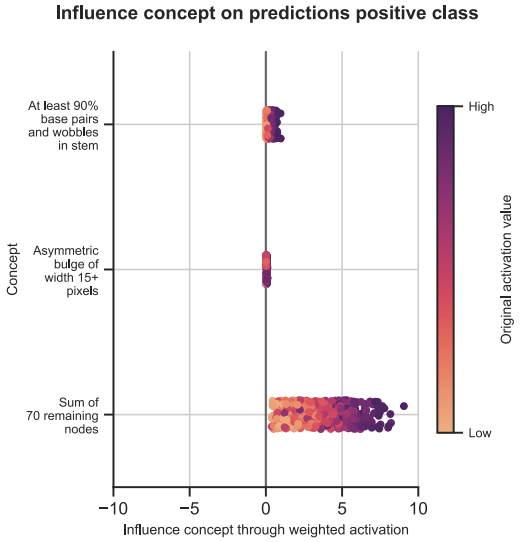

Figure C.11: Influence values of the nodes aligned with concepts and of the sum of all remaining nodes in the CW layer on the positively classified pre-miRNAs. The influence is defined by the weighed activation value. The color of the values is based on their original activation value.

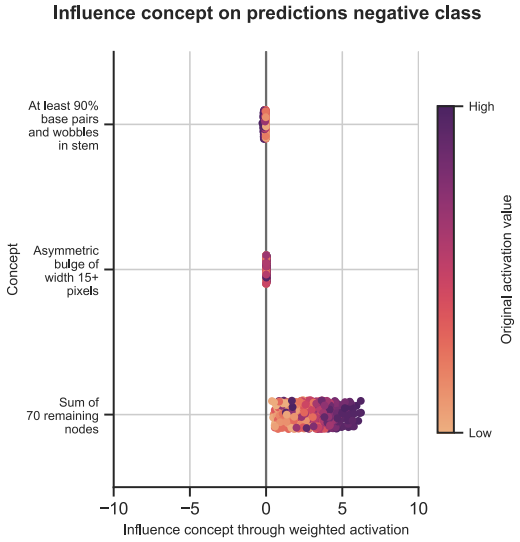

Figure C.12: Influence values of the nodes aligned with concepts and of the sum of all remaining nodes in the CW layer on the negatively classified pre-miRNAs. The influence is defined by their weighed activation value. The color of the values is based on their original activation value.

Tree-based Concept Importance Explanation Another way to explain the importance of concepts for the pre-miRNA predictions on a global level would be using a decision tree, which can provide non-linear explanations. More specifically, this tree can generate a hierarchical explanation based on concept importance results that holds for the complete dataset. Consider an example shown in Figure C.13, where we illustrate this type of explanation using the *Large asymmetric bulge* (concept 4) and *At least* 90% *base pairs and wobbles in stem* concepts (concept 2) based on their influence results on the predictions. Since we normalized the class weights of the two output nodes before calculating the influence values shown in the tree, we refer to the values as the *normalized influence values*. We normalized the weights since the weights of one output node can be much larger than the other, resulting in a situation where the influence values are so diverging that the explanations may become misleading. The normalization is applied to the two weight vectors separately. Hence, all values in both are scaled onto a range of $[0, 1]$ using the minimum

and maximum weight value in the vector of interest.

From the tree in Figure C.13, we can conclude that the model has classified 37.5% of the pre-miRNA instances as *positive* due to a larger influence of learned concept 2, combined with a smaller one of learned concept 4. For most instances classified as *negative*, we find that concept 2 has a smaller influence and 4 a larger one. This is in line with the concept importance results given in 6.2.2. As of now, the tree is not very informative since only two concepts are included. However, this type of explanation could potentially become more useful if more are added.

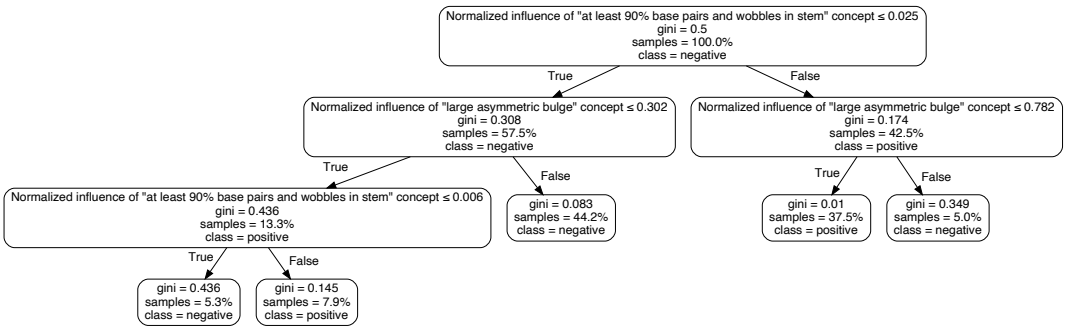

Figure C.13: Concept importance explanation given with a decision tree based on the normalized influence results of the *Large asymmetric bulge* and *At least* 90% *base pairs and wobbles in stem* concepts.

C.3 Model with 1 output node

As concluded from the domain expert evaluation and discussion on the concept importance explanations (sections 6.3 and 7.2), the understandability of our explanations is limited due to a design choice of our CNN. Namely, we would prefer to have only one output node instead of two, with one for each pre-miRNA class. In the case of one output node, we would use a threshold that specifies for the output values to which class the input instance mapped to the output would obtain.

We conducted some experiments with this type of model. We changed the Softmax activation function of our final FC layer into the Sigmoid one and the loss function from cross-entropy to binary cross-entropy. The training and evaluation of this model before converting the last BN layer into a CW one was straightforward. We obtained accuracy scores above 90%, meaning the performance is on par with the model having two output nodes. However, the training of this new CNN **after** including the CW layer was very difficult. We show this with the results of one of our experiments, where we used only the *Large asymmetric bulge* concept (concept 4) and a learning rate of $1.0e-5$. As shown in Figure C.14, the accuracy results per epoch obtained on the training folds with this model look more or less promising, as the scores are increasing but with some serious instabilities. The scores obtained on the validation folds, shown in Figure C.15, illustrate a very different situation. The lines composed of the accuracy scores are very unstable and show that the scores decrease extremely after increasing for some epochs every 5-10 epochs.

This unstable behavior during training also occurred using different learning rates, different batch sizes, and other concepts defined in section 5.3.1. This made us conclude that training a CW model with only one output node in the final FC layer is not straightforward. Forcing the model to use the output of the CW layer, which are the whitened data representations summarized into scalars, to immediately generate class predictions may be too rigid. However, more efforts should be made into training this model, as the output will certainly increase the understandability of the concept importance explanations.

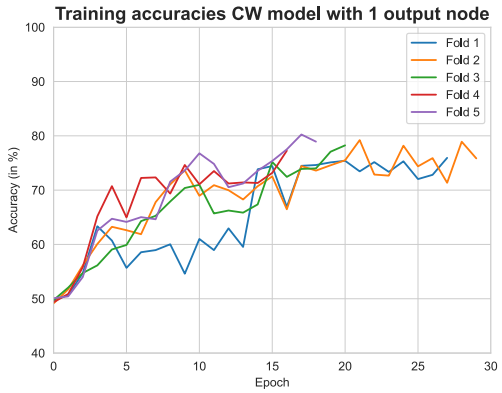

Figure C.14: Accuracy per epoch using the CW model with one output node and the five training *modhsa*-data folds.

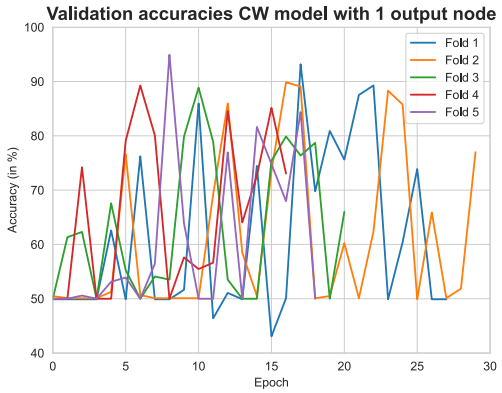

Figure C.15: Accuracy per epoch using the CW model with one output node and the five validation *modhsa*-data folds.

C.4 Definition of novel pre-miRNA concepts

In this section, we provide more examples of the images highly activated by the most influential nodes in the CW layer handling residual image information (i.e., they are not aligned with a pre-defined concept). First, we do this for the most influential node for the *positive* pre-miRNA class, which is the 36^{th} node in the layer. In Figure C.16, some of the most highly activated test set images (besides those in Figure 6.24) are given. These images also show this node is focusing mostly on green-yellow pixel pairs, denoting A-U nucleotide pairs, in the stem area. Again, this leads to the conclusion that the model seems to find a sequence of A-U pairs in the stem area of pre-miRNAs important for classifying the sequences as *positive* pre-miRNA. We find the node also focuses on relatively high blue-colored bars, being part of a bulge. And besides A-U pairs, the node also focuses on the other type of base pairs, namely the G-C pairs.

Next, we provide more highly activated images for the node in the CW layer most influential for the *negative* pre-miRNA class, which is the 4^{th} node, in Figure C.17. We see all images contain several large asymmetric bulges. Some of the receptive fields highlight different regions in different images. However, most seem to highlight black-colored pixels, representing gaps. These gaps are part of asymmetric bulges, as shown in the images. Also, the highlighted regions seem to illustrate the node's focus on the whitespace to the right of the sequence. In section 6.2.2, we concluded the focus of the node is more on the central part of the sequence. However, this does not necessarily hold for the images shown here. This once again illustrates the difficulty that comes with deriving meanings on the CW model's learning of information relevant for the pre-miRNA detection task based on diverging receptive fields.

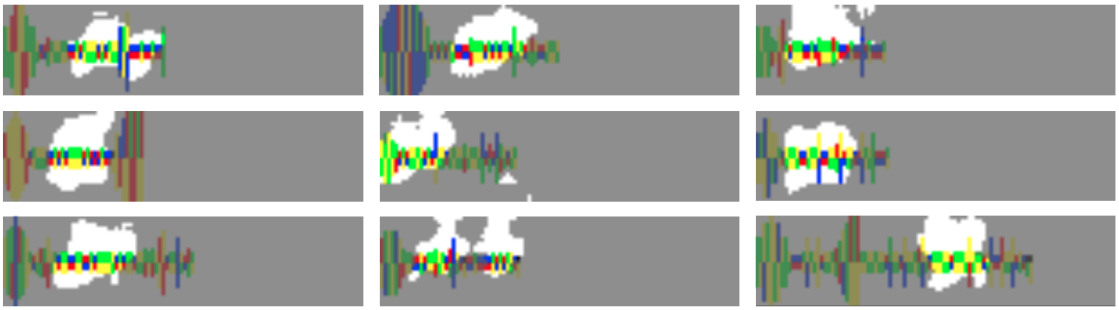

Figure C.16: Highly activated test set images for the most influential node in the CW layer for the *positive* pre-miRNA class (i.e., node 36).

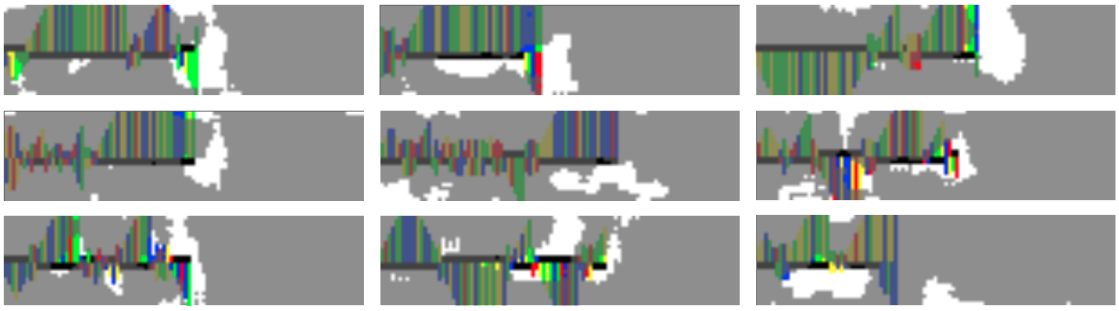

Figure C.17: Highly activated test set images for the most influential node in the CW layer for the *negative* pre-miRNA class (i.e., node 4).

C.5 Evaluation with Domain Expert

In this section, we provide the questions used to interview the domain expert for evaluating the proposed framework and results. As explained in section 6.3, evaluations of interpretability methods indirectly evaluate the ML system's or model's motivation for interpretability. In our case, this motivation is *to promote knowledge discovery*. Also, we set our focus on the *understandability* goal ML interpretability, since CW's intrinsic interpretability should ensure the *fidelity* goal is met. With this motivation and goal, we defined that the main goal of the evaluation should be to assess the usefulness of the proposed framework and results. This usefulness is determined by letting the domain expert determine whether the generated concept-based explanations are understandable and whether they may assist in finding structural pre-miRNA characteristics.

We evaluated four different aspects of the framework and results, namely the provided *interpretation*, the generation of *concepts*, the concept whitening and explanation generation *approach* and the framework in *general*. Following are the different questions used to evaluate these aspects of the framework together with the domain expert.

Interpretation

- What do you conclude from the explanations given by the CW model?
  - What do you think of the information learned from the base pairs and wobbles in stem concept?
    - Is it in line what you expected?
    - Are the explanations probable, i.e. are they in line with your knowledge on this potentially discriminatory feature of (pre-)miRNAs?

    - Is the definition of the concept valid, so is it in line with your mental model of the concept?
  - What do you think of the information learned from the large asymmetric bulge concept?
    - Is it in line what you expected?
    - Are the explanations probable, i.e. are they in line with your knowledge on this potentially discriminatory feature of (pre-)miRNAs?
    - Is the definition of the concept valid, so is it in line with your mental model of the concept?

- In terms of the handling of residual information, would you prefer to know what all remaining nodes might have learned or do you prefer how it is done currently (only the most influential nodes are explained)?
- In your opinion, are we missing any concepts potentially relevant for the pre-miRNA classification task solved with this concept-based interpretability method?

Concepts

- Are the concept representations understandable? Would they be understandable for a domain expert that has not seen the encoded pre-miRNA sequences yet?
- Are the definitions of the current concepts clear and understandable?
- Would a domain expert be able to define new concepts for the model?

Approach

- Is it clear what the goal of the concept whitening procedure is? Is it clear why this is relevant/useful?
- Is it clear how the concept importance explanations are created (e.g., two output nodes, weighting procedure)?

General

- Overall, would this framework be useful in practice (given that one explains how to define and annotate concepts and how to train it properly?)
- Do you think the explanations can support understanding and deriving structural characteristics of pre-miRNAs from the trained model?