Learning hierarchical motif embeddings for protein engineering

by

Thrasyvoulos Karydis
Dipl., University of Patras (2014)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

Signature redacted

Author .....
Program in Media Arts and Sciences
School of Architecture and Planning
December 15th, 2016

Signature redacted

Certified by .....
Joseph M. Jacobson
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by .....
Pattie Maes
Academic Head
Program in Media Arts and Sciences

Learning hierarchical motif embeddings for protein engineering

by

Thrasyvoulos Karydis

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on December 15th, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences
at the
Massachusetts Institute of Technology

Abstract

This thesis lays the foundation for an integrated machine learning framework for the evolutionary analysis, search and design of proteins, based on a hierarchical decomposition of proteins into a set of functional *motif embeddings*. We introduce, **CoMET - Co**nvolutional **M**otif **E**mbeddings **T**ool, a machine learning framework that allows the automated extraction of nonlinear motif representations from large sets of protein sequences. At the core of CoMET, lies a Deep Convolutional Neural Network, trained to learn a basis set of *motif embeddings* by minimizing any desired objective function.

CoMET is successfully trained to extract all known motifs across Transcription Factors and CRISPR Associated proteins, without requiring any prior knowledge about the nature of the motifs or their distribution. We demonstrate that motif embeddings can model efficiently inter- and intra- family relationships. Furthermore, we provide novel protein meta-family clusters, formed by taking into account a hierarchical conserved motif phylogeny for each protein instead of a single ultra-conserved region.

Lastly, we investigate the generative ability of CoMET and develop computational methods that allow the directed evolution of proteins towards altered or novel functions. We trained a highly accurate predictive model on the DNA recognition code of the Type II restriction enzymes. Based on the promising prediction results, we used the trained models to generate de novo restriction enzymes and paved the way towards the computational design of a restriction enzyme that will cut a given arbitrary DNA sequence with high precision.

Thesis Supervisor: Joseph M. Jacobson
Title: Associate Professor of Media Arts and Sciences

Learning hierarchical motif embeddings

for protein engineering

by

Thrasyvoulos Karydis

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on December 15th, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences
at the
Massachusetts Institute of Technology

The following people served as readers for the thesis:

Academic Advisor ... Signature redacted .....

Joseph M. Jacobson

Associate Professor of Media Arts and Sciences, MIT

Signature redacted

Reader .....

Neil Gershenfeld

Professor of Media Arts and Sciences, MIT

Signature redacted

Reader ....

Manolis Kellis

Professor of Computer Science, MIT

Acknowledgments

This thesis could not have been completed without support from many amazing people.

The mentors

Firstly, I would like to thank my advisor, Joe Jacobson, for introducing me to some of the most challenging problems in biology, and for his constant support and excitement for the work of this thesis.

I'd like to thank Neil Gershenfeld, for giving me the opportunity to conduct research across disciplines, the knowledge and the resources to make almost anything and the privilege to meet some of my heroes in science and engineering.

I would like to thank Manolis Kellis, for his mentorship and support of this thesis, but most importantly for welcoming me to work with him and his students.

Last but not least, I would like to thank Andreas Mershin, for the unforgettable experiences we shared the past three years at MIT and around the globe. His boldness and enthusiasm for scientific research is a continual inspiration. I am grateful for teaching me how to ask questions, but also how to question answers.

The ~~lab-mates~~ friends

I would like to thank Charles, Lisa, Pranam, Kfir, Frank and Allan for offering the best working environment someone could wish for. I am especially grateful to Noah, for the long nights and deep conversations, without which this thesis would not have been possible; oh, and for his appreciation of my jokes. I'd like to thank Eric, James, Grace, Ben, Sam, Amanda, Will, Nadya, Prashant, Michael, Rui for accepting me in the CBA family; they are constantly inspiring me with their work. Special thanks to Niko for all the time we spent inside and outside the lab. Lastly, I'd like to thank everyone at the Media Lab for making it a truly unique place to be part of.

The "people behind the scenes"

Many thanks to Joe and Ryan, for making ordering things "happy". I'd also like to thank John and Tom for teaching me how to use the machines and making sure we are all safe at the shop.

The "best admins in the whole universe"

Jamie, Keira and Linda – I have no words; sincerely, thank you.

The family

I would like to thank Katerina, for her love and kindness to guide me through all the hard decisions in my academic life; Anastasia, for putting up with my terrible work/life balance and for making my life more beautiful and less boring; All my buddies, who make it feel like I never left home, even if I see some of them once a year; my "older brother" Konstantinos, whose life constantly motivates me to be better. Finally, I have the utmost gratitude and appreciation for my parents and brother, for their unconditional love, constant support and encouragement.

Special thanks

To Aditya Khosla, for his invaluable support during the development of this work.

To David Kong, for creating EMW and for striving to make great things happen.

To my friends from EMW Streetbio.

To all the students and faculty from How to Grow almost Anything.

To GlaxoSmithKline, for their support of this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Biological engineering is not necessarily understanding systems but rather, I want to be able to design and build biological systems to perform particular applications.*
>
> —Drew Endy, *2005*

Technological advances in the past decade have allowed us to take a close look at the proteomes[1] of living organisms. As a result, more than 120,000 protein structures have been solved and are readily available (source: `www.rcsb.org`) to the public, a number that grows exponentially. Looking through the protein structures, one comes across a vast diversity of molecular machinery, involving intricate mechanisms that perform key cell functions, such as genome replication, energy production or immunity (Fig. 1-1). Today, a large part of scientists working in the field of synthetic biology, are trying to formulate rules and principles behind the construction of such proteins, a methodology known as *rational protein design*, which resembles the process behind the making of a complex mechanical or electronic device.

Biological cells, though, do not have a degree in engineering. Cells are able to adapt to changes in their environment, such as introduction of new chemicals or attacks from new viruses, through a sophisticated encoding and an evolutionary search algorithm. The efficiency of this method lies in the ability to generate millions of different alternatives, immediately put to test in distinct copies of the organism,

[1] A proteome is the set of proteins expressed by a genome, cell, tissue, or organism at a certain time.

a process known as natural selection. By looking at the proteomes of the current living organisms, we are essentially taking snapshots of the successful results in this evolutionary process of continuous adaptation to the environment. Could we leverage the available information to design new proteins, without the need for millions of years of Darwinian evolution?

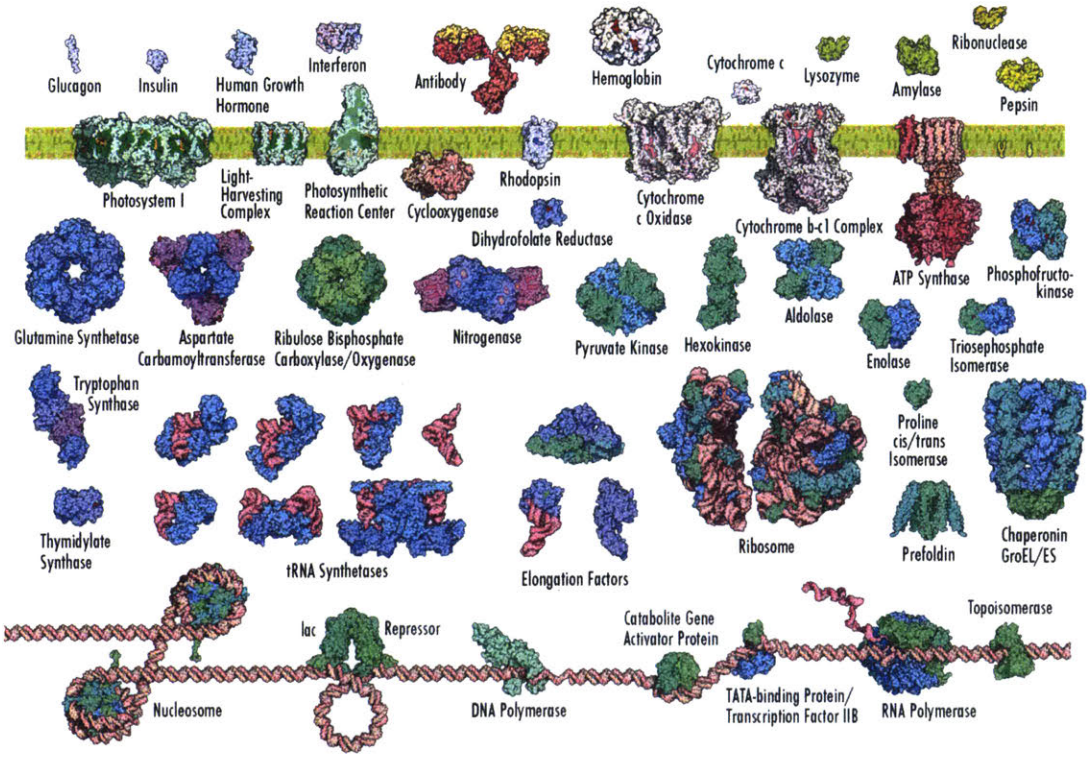

Figure 1-1: Atlas of protein structures involved in key biological processes. Source *ProteinDataBank*. Image credit: David Goodsell.

To faithfully answer this question, we have to take into account that design was not a big component in this evolutionary process. Thus, by employing engineering methodologies, we can do much better than random combinatorics of single amino acids. In fact, by looking carefully at the sequences and structures of proteins across species, we can find patterns of similar amino acid composition that have distinct functions within the larger protein structure. These patterns, or *motifs*, are a mixture of secondary and tertiary structure elements, that are combined together to form the final protein with complex functionality. We can further parallelize this structural and functional modularity with similar concepts from electromechanical system

engineering (see Figure 1-2).

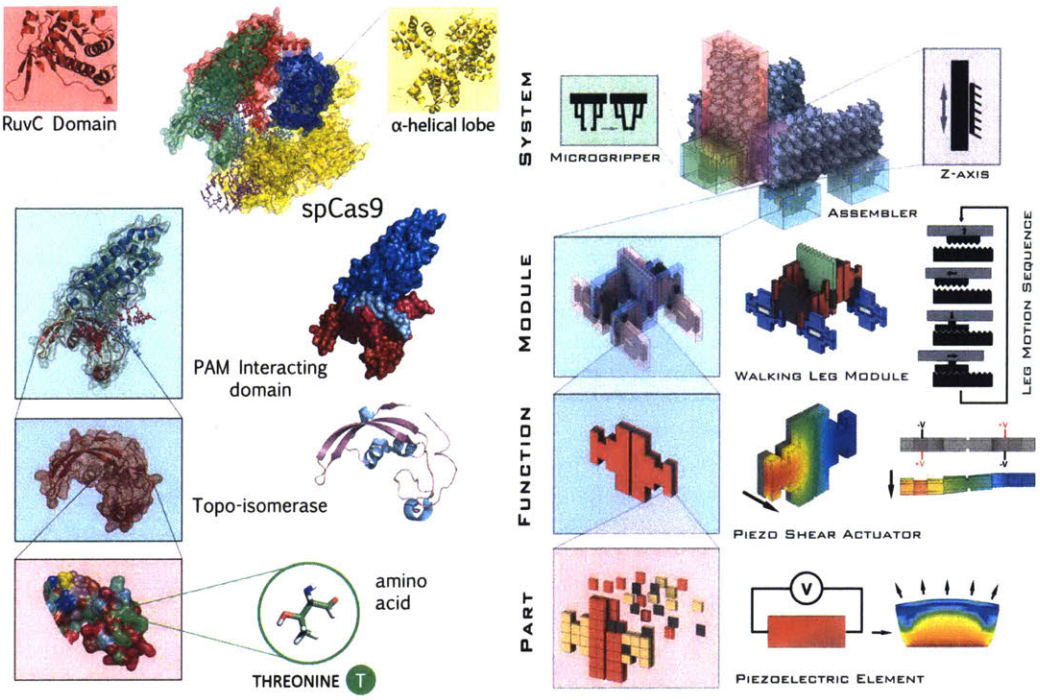

Figure 1-2: (left): Hierarchical decomposition of protein *spCas9*. Initial structure downloaded from *RCSB* (4UN3) and visualized with *PyMOL*. (right): The various levels of hierarchy in the design of a modular machine, by Will Langford, MIT Center for Bits and Atoms.

Thus, from an engineering perspective, we pose ourselves the following questions:

- Is biology building its arsenal of proteins using a hierarchical set of building blocks - *motifs* - on top of the primary amino acid code?
- If so, could we leverage the available information about the sequence. structure and function of proteins found in nature at this moment, in order to learn those motifs?
- Lastly, given the basis set of motifs at a particular level of representation, how can we use them to engineer proteins with altered or novel functions?

This thesis lays the foundation of an integrated machine learning framework for the evolutionary analysis, search and design of proteins, based on a hierarchical de-

composition of proteins into a set of functional motif embeddings, utilizing the latest advances of Deep Convolutional Networks (DCN). Current work focuses on the motif embeddings representation per se, paving the path towards homology based search and ultimately generative modeling for design.

We introduce, **CoMET** - **Co**nvolutional **M**otif **E**mbeddings **T**ool, a machine learning tool that allows the automated extraction of nonlinear motif representations from large sets of protein sequences. At the core of CoMET, lies a Deep Convolutional Encoder, that is trained to learn a basis set of motif embeddings by minimizing a set of objective function. The learned protein embeddings can be visualized as one-dimension sequence logos, or as points in a multi-dimensional embedding space.

We show that using the embeddings you can search for evolutionary distant homologous sequences and we define two metrics to compare the search results with current linear sequence motif based methods. CoMET departs from the motif extraction status quo, namely multiple sequence alignment algorithms, which require fine hand tuning of motif search parameters, and performs a search automatically tuned to produce results satisfying a given optimization criterion, such as maximum functional or structural similarity. We subsequently use CoMET to reconstruct phylogenetic trees of protein families. Lastly, we investigate the generative ability of CoMET and develop computational methods that allow the directed evolution of proteins towards altered or novel functions.

1.1 Motivation

Proteins perform the vast majority of functions responsible for our health and well-being. Thus, designing unique proteins or engineering existing ones is paramount for the implementation of safe and successful therapeutics. To conclude the introduction, we present to the reader two of our strongest motivational drivers behind this work, namely gene therapy and drug discovery.

Drug Design Drugs are, in majority, organic small molecules purposed to activate or inhibit a specific function of a multifunctional protein or disrupt proteinprotein interactions, resulting in the treatment of a disease [1]. Despite the paramount advances in many of the scientific, technological and managerial factors that should tend to raise the efficiency of commercial drug research and development, the number of new drugs approved per billion US dollars spent on R&D is steadily falling [2]. In the heart of this declining curve, is our inability to understand the complex mechanisms involved in the process of biomolecular recognition between a protein and its epitope.

Biomolecular recognition drives fundamental biological functions in any living organism, including gene regulation and sensing environmental stimuli [3]. It involves the interaction of a protein, usually a receptor or enzyme, with molecules present in living organisms, ranging from parts of macromolecules such as pieces of DNA to small molecules such as neurotransmitters and odorants. Efficient biomolecular recognition requires exquisite control of affinity and specificity [4].

In the lack of that control, conventional drug discovery is performed by screening the target proteins through large molecular assays (typically in the order of a million molecules) to identify molecules that or activate their function. These screenings are usually informed by structure based strategies [5] and in some times by preceded by computational screening [6].

Unfortunately, through the above process, while it is possible to find molecules that work, there is no information gained about their mechanism of action. Thus, every new target requires the same amount of time and effort spent to screen through the entire molecular library, as if we were to discover a drug for the very first time. On the other hand, computational protein design when coupled with machine learning algorithms results in a net gain of information from each drug discovery, which in turn makes the discovery of the next one faster and easier.

Gene Therapy Next Generation Sequencing (NGS) technologies are providing a constant influx of genomic data while computational biologists are catching up inventing new algorithms with the capacity to extract meaningful information out of

the big data. This fast-paced progress has led to astounding discoveries, a large class of which is the causal connection of single-point mutations in our genetic code to particular illnesses, including obesity [7], sickle-cell anaemia [8] and assorted neurological conditions. Thus, gene therapeutics, i.e. precisely editing specific bits of the human genome will catalyze the next generation of medical applications.

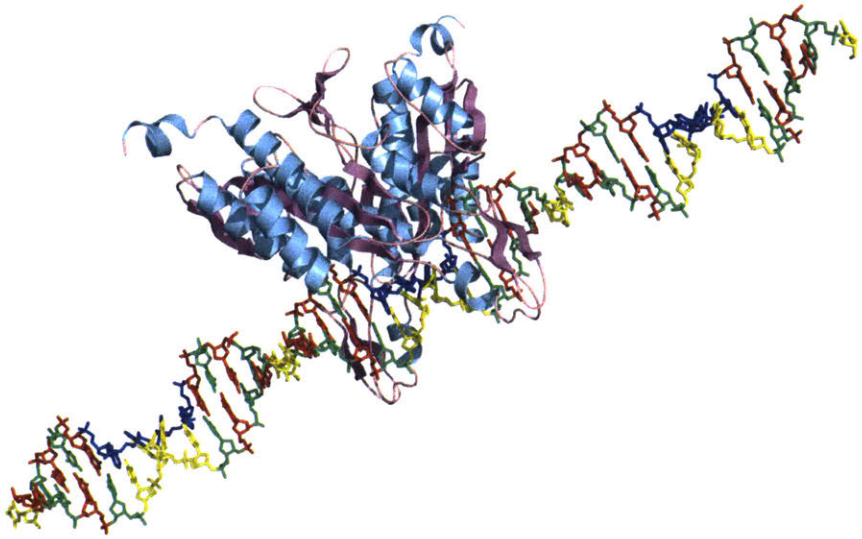

Figure 1-3: *EcoRI* protein in complex with the target DNA sequence. Structure downloaded from *RCSB* (1ERI) and visualized with *PyMOL*. The protein is colored by secondary structure and the DNA by the four base code.

Yet, current genome editing tools show significant off-target activity (ZFNs, TALENs, CRISPR) leading to spurious cuts at sites different than the one targeted [9, 10]. These complications render it possible for a gene therapy to cause unwanted gene mutations that will lead to possibly cancerous malfunction or death of the cell. While there have been many attempts to increase the specificity [11] and sequence space [12] of current tools, those usually increase significantly the complexity of experimental protocols without reaching sufficiently low frequency of off-target cleavage.

On the other hand, nature-engineered DNA cutting enzymes called *endonucleases* found in bacteria and archea, have very high sequence specificity with minimal off-target cleavage, which renders them the ideal gene editing tools. Endonucleases, also known as restriction enzymes, have two active sites: a recognition site which binds

to specific DNA sequences and a cleavage site, which cuts the double stranded DNA (Figure 1-3). Nevertheless, endonucleases have a significant shortcoming, i.e. we are currently limited to those found in nature with no general means for altering their recognition sites in order cut an arbitrary DNA sequence of interest.

By developing computational methods for the programmable design of the active sites of endonucleases, we are providing to the field of gene therapeutics a novel gamut of gene editing tools with enhanced specificity, that can eventually lead to safer in vivo applications.

Chapter 2

Functional decomposition of proteins into hierarchical motif embeddings

Functional annotation, taxonomy and search of proteins, today, is fundamentally based on the discovery of short sequence patterns or *motifs*. Yet, as the number of available protein sequences increases exponentially, existing motif discovery tools and practices, based on manual identification or probabilistic sampling, are reaching their limits in required time and effort. In addition, due to the extreme segregation of proteins into thousands of highly specific families [13], there is a plethora of uncharacterized proteins with unknown functions. To overcome these limitations, we introduce *motif embeddings*, a compact real-valued vector produced from a hierarchical decomposition of protein sequences into motifs. To learn the embeddings, we employed state-of-the-art deep learning techniques and implemented (**CoMET**) (**Co**nvolutional **M**otif **E**mbeddings **T**ool), an automated program for the extraction of protein motif embeddings from arbitrarily large protein sequence datasets.

At the core of CoMET is a Deep Convolutional Network Encoder, which upon training learns a hierarchical motif representation from a set of input protein sequences. Here, we introduce the nuts and bolts of CoMET and evaluate the performance of motif embeddings in a series of real biological applications. Firstly, we show

that CoMET extracts all known motifs across major protein super-families, when trained on large protein datasets without requiring any prior knowledge about the nature of the motifs or their distribution. Concurrently, we apply unsupervised learning techniques to identify a minimal number of functional clusters that can explain the sequence variance within a super-family, and compare the results with the existing family categorization. Lastly, we employ CoMET on $C2H2$ Zinc Finger proteins, in the quest to find a minimal hierarchical motif decomposition, that will serve as building blocks for rational protein design.

2.1 Background and Related Work

Nowadays, large protein datasets become available to the scientific community in a daily basis through Next Generation Sequencing (*NGS*) driven bio-technologies. Uniprot[1] (`http://www.uniprot.org/`), the largest online, publicly-available protein database, houses more than *555,000* protein sequences with experimental evidence (e.g. through RNA-Seq [14]) and more than *63 million* translated from coding regions of sequenced genomes ((Fig. 2-1). Discovery of representative motifs is an essential step in the structural analysis [15], functional annotation [16] and taxonomy [17] of those protein sequences.

Publicly available motif databases, including **Pfam** [13], eukaryotic linear motif database - **ELM** [18], **Prosite** [19] and **ScanSite** [20] contain a comprehensive list of sequence motifs and domains in the form of regular expressions or profile HMMs. The largest of them, Pfam, comprises of more than 16000 (as of version 30.0), manually curated entries of well-documented protein motifs.

Largely, the motifs are discovered by manually identifying protein regions and residues that form known functional groups, such as enzyme catalytic sites or metal ion binding amino acids. The databases collect the motifs from studies in certain protein families (i.e. transcription factors, enzymes) by means of a multiple sequence alignment of select proteins, and subsequently enrich them by searching for occur-

[1]Data as of release 2016_07.

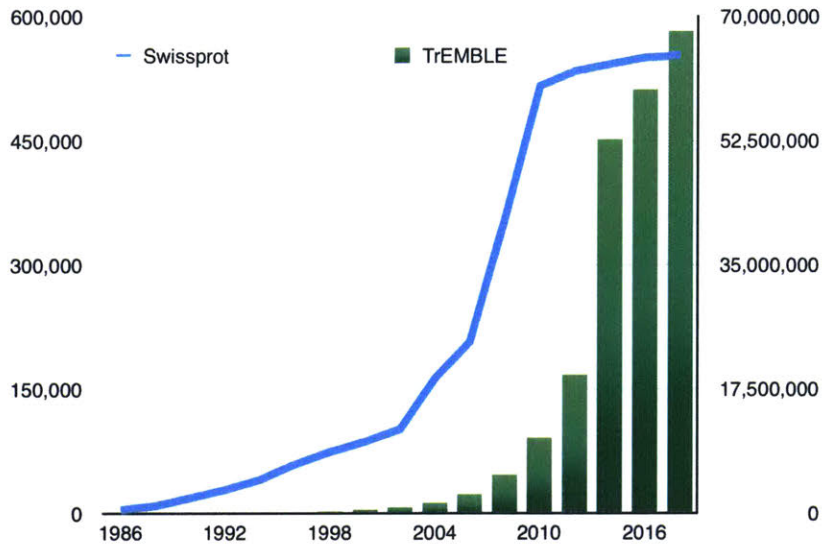

Figure 2-1: Number of experimentally validated [Swissprot] and translated from genomic data [TreMBLE] protein sequences increased exponentially after the coming of NGS bio-technologies in the past decade.

rences of those patterns in protein sequence databases.

Obviously, these resources cannot be utilized for the discovery of an unannotated sequence motif that is suspected to be a common feature in a given protein set. In other words, running a motif discovery methodology fueled from current knowledge, renders it impossible to discover protein motifs that correspond to novel, unknown functions or structures. Furthermore, manual identification of important protein sites and residues is quite challenging when the structure has not been solved, which is the case for more than 80% of experimentally validated protein sequences. Lastly, the motifs in the databases are largely overlapping, as they are maintained with the purpose of an exhaustive segregation of proteins to families of identical sequences. The latter inhibits the use of those motifs as building blocks for modular protein design.

A semi-automated approach in de-novo motif discovery, originating from the early days of computational biology, is the use of stochastic search (Gibbs sampling) [21] or probabilistic models [22]. Two of the most prominent, publicly-available protein[2]

[2]The tools are used also for nucleotide sequences but this escapes the scope of this thesis. Moreover, there exists an interesting integration of multiple DNA motif discovery tools into a single automated interface: `http://fraenkel.mit.edu/webmotifs.html`.

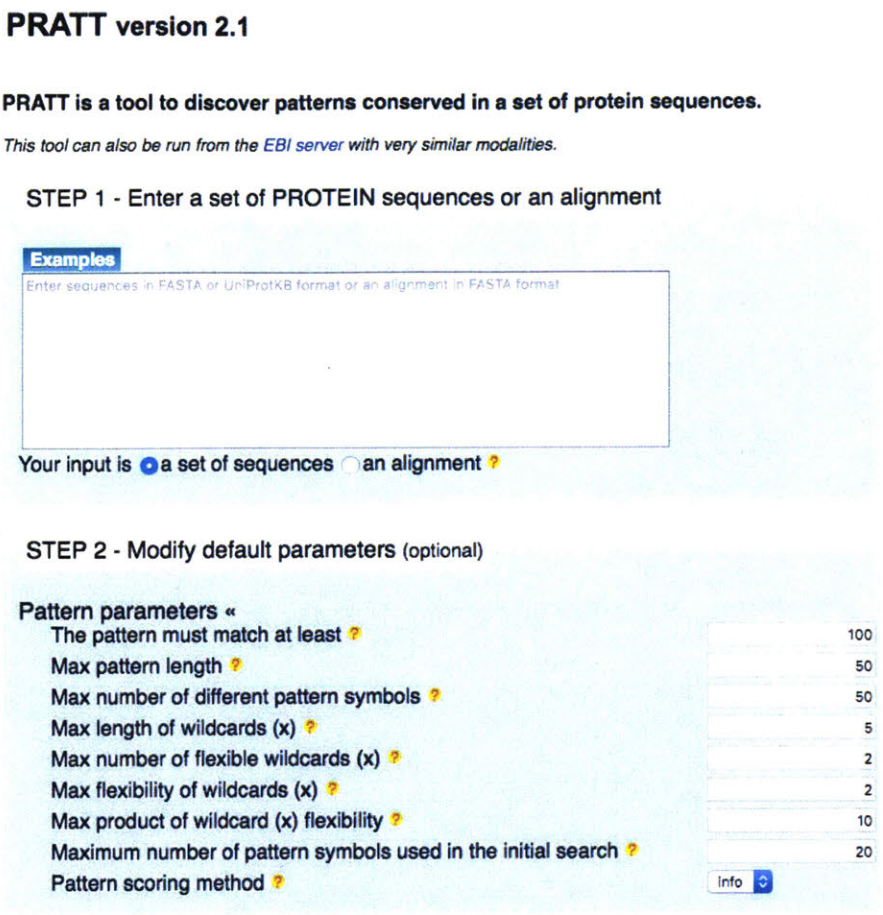

Figure 2-2: A contemporary motif extraction tool, which is semi-automated as it requires user input for various motif parameters.

motif extraction tools using these methods are *MEME* (`http://meme-suite.org/tools/meme` and *PRATT* (`http://web.expasy.org/pratt/`.

These tools, while not requiring any prior knowledge about the proteins to extract sequence motifs, they come with a set of inherent limitations. Notably, you have to know a priori the statistics of the motifs, such as the length and the size (see Figure 2-2). Also, algorithmic complexity results into intractable run-times and stochastic sampling produces poor results for large size datasets (e.g. the *MEME* tool allows inputs only up to 1000 sequences). Last but not least, large protein domains, such as those contained in Pfam, can be defined quickly from alignments of evolutionarily related sequences, but the identification of short sequence motifs, potentially shared between proteins that appear evolutionarily unrelated, is much harder.

To overcome the above limitations, we developed **CoMET** (**Co**nvolutional Motif **E**mbeddings **T**ool), a computational tool for the hierarchical decomposition of protein sequences into a set of motifs. **CoMET** maps every protein into a binary vector called *motif embedding*, which holds information about the presence of any non-linear combination of a fixed set of protein motifs. At the heart of this work, lies an adaptation of *Deep Convolutional Neural Networks* (DCNNs), which we train to learn the underlying sequence patterns for each motif embedding. DCNs, have been proven extremely successful in image, video and speech classification and recognition tasks [23, 24], as well as in computational genomics [25].

Most importantly, the past few years novel methods have been developed [26, 27], that allow us to understand the internal feature representations that the networks learn, as well as, visualize them in the input space. Novel generative modeling techniques allow the use of trained DCN to produce new, unseen data with desired properties specified as the network outputs [28, 29]. The latter, is the strongest motivational driver for our choice of network architecture, as the ultimate use case of motif embeddings is rational protein design.

Lastly, there exists a variety of methods that embed proteins in distributed representations, with the most notable ones being ProtVec [30] and ProtEmbed [31], demonstrating successful completion of a series of family and structure classification tasks. While we share similar aspirations for the applications of protein embeddings with the authors of aforementioned articles, this work differentiates itself through three major ways: a) introducing a hierarchical decomposition architecture to map the protein sequences to embeddings; b) allowing the direct interpretation of the components of an embedding vector into a set of protein sequence motifs (hence the name *motif embeddings*); c) the Convolutional Network architecture allows diversity in the information used as input (secondary structure, conservation etc.) by simply having extra "channels" in a common convolutional layer.

2.2 CoMET: Convolutional Motif Embedding Tool

In this section, we will describe the core architecture of **CoMET**, as well as the network architectures we used to solve supervised and unsupervised learning tasks using the motif embeddings. We will define a set of metrics to be used as training objectives for training, as well as for the evaluation of learned motif embeddings. Lastly, we will discuss the rationale behind searching and selecting for the optimal hyper-parameters.

2.2.1 Core Architecture

The core architecture of **CoMET** is a Convolutional Network (ConvNet) with rectification, followed by a Max-Pooling (gMP) stage and a set of Fully Connected (FC) layers (Figure 2-3). The convolution stages have trainable motif detectors D and rectification thresholds b, while the fully-connected network stages have trainable weights W; the max-pooling stage has no trainable parameters. Each computational stage is further explained in the following subsections.

Essentially, a **CoMET** model takes a single (protein) sequence $S = (s_1, s_2, \ldots, s_L)$ with alphabet[3] $A = \{A, B, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, X, Y, W, Z\}$, and produces a real-valued vector (motif embedding) $\mathcal{M}(S) = (m_1, m_2, \ldots, m_N)$, where N is the number of filters used in the convolutional network. The motif embedding $\mathcal{M}(S)$ for sequence S is computed by a feed-forward pass through the **CoMET** encoder, starting with convolution and ending in a fully-connected neural network. Symbolically,

$$\mathcal{M}(S) = FC_W\Big(gMP\big(ConvNet_M(S)\big)\Big) \tag{2.1}$$

The specific architecture of a given **CoMET** (number of motif detectors; deep versus shallow ConvNet; deep versus shallow Fully-Connected network) will differ based on the properties of the input dataset as well as the upstream network architecture for each application (see subsection 2.2.2. Yet, the general form is the one described

[3]We allow single letter ambiguity codes (B, Z) as well as the letter X for total ambiguity.

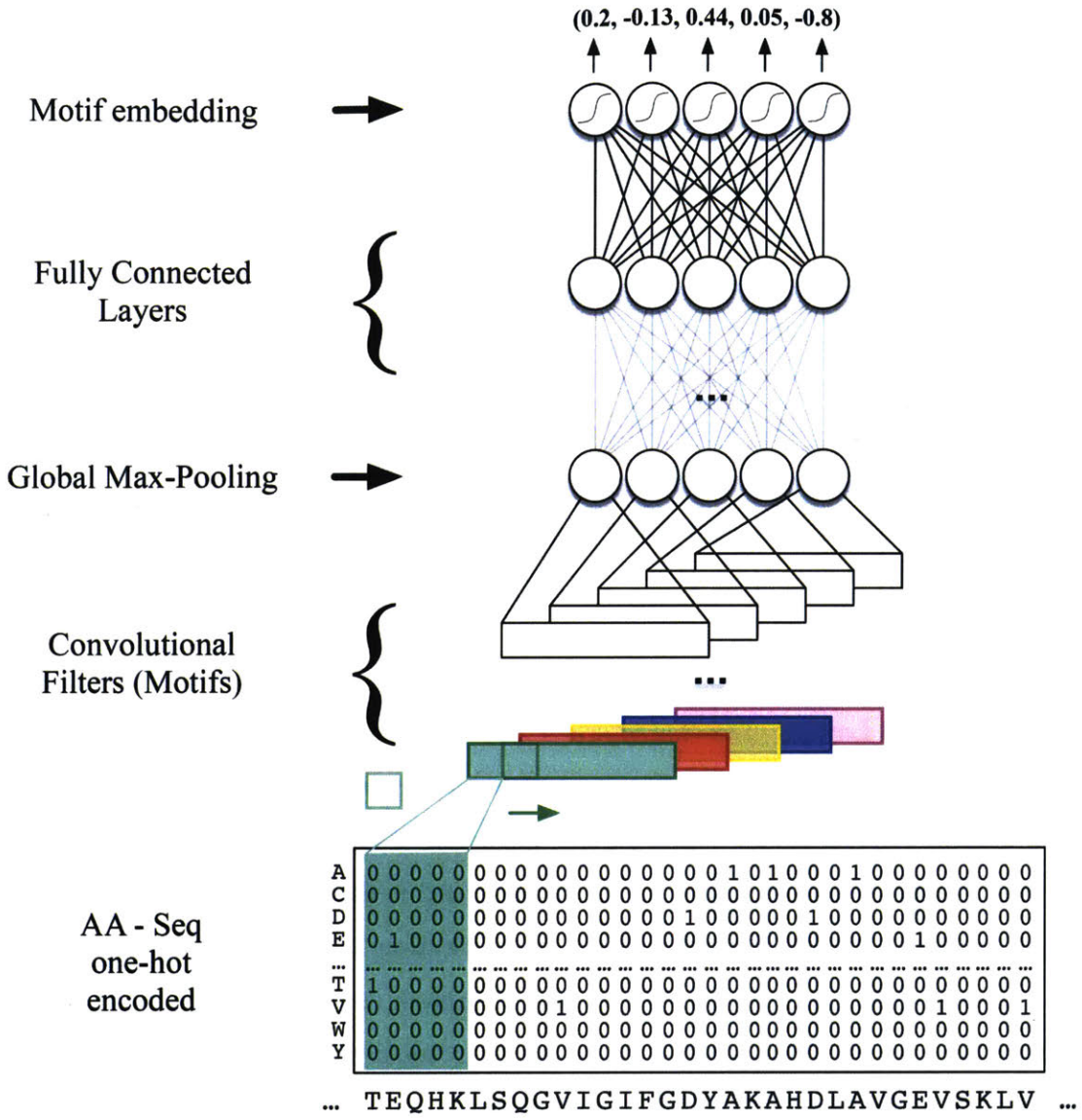

Figure 2-3: The core architecture of **CoMET**. It encodes a protein sequence of arbitrary length to a fixed length real-valued vector we call *motif embedding*.

here and remains the same in all instances. Analogously, while the choice of a loss function depends heavily on the particular training objective, the upstream gradients are propagated through the **CoMET** core architecture in order to train the weights and motifs thereof.

The rest of this section provides further details about the implementation of each computational stage of the **CoMET** encoding architecture.

Input Encoding **CoMET** accepts as input a protein sequence of single-letter amino acid codes, of arbitrary length. We begin by transforming the protein sequence to a "1-of-20" encoding (or "one-hot" encoding). Specifically, a given protein sequence $S = (s_1, s_2, \ldots, s_L)$ of length L is transformed into an $L \times 20$ array, where each column has zero values in all rows, apart from, typically, the row corresponding to the amino acid at the specific position. For a given amino acid code s_i, its "one-hot" encoding is:

$$OH(s_{ij}) = \begin{cases} 1 & \text{if } s_i \text{ is the } j\text{-th element of the alphabet } A - B, Z, X \\ 0.5 & \text{if } s_i \text{ is } B \text{ and } j \text{ is the order of the element } D \text{ or } N \\ 0.5 & \text{if } s_i \text{ is } Z \text{ and } j \text{ is the order of the element } E \text{ or } Q \\ \frac{1}{20} & \text{if } s_i = X \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Convolutional Layers The characteristic feature of a **CoMET** architecture is that it comprises a set of stacked $1 - D$ convolutional layers, where the input is convolved with several tunable arrays called *filters*. Selection of the number of filters as well as filter length, depends highly on the dataset and is discussed thoroughly on the hyper-parameter optimization discussion below (Section 2.2.4). As noted in the work of Alipanahi et al. [25], here in the context of protein sequences, a one-dimensional convolution over the 20-channel input resembles a "motif scan" operation in a PWM- or PSAM- based motif extraction model. In contrast with a PWM/PSM though, the filters matrices of the convolutional layer have unconstrained norms and values.

The filters act upon the entire sequence (full convolution) and there is no dimensionality reduction between each convolutional layer. All convolutional layers apply a *rectified linear unit* (ReLU) [32] to the sum of the computed convolution values with an activation threshold b. Thus, if the convolution of a filter in a specific position of the sequence is less than the threshold, the resulting score for that position is set to zero.

For an input protein sequence of length L, the output of the convolutional stage is an $L \times N$ array, where N is the number of filters (all of the same length m) for

a given **CoMET** model. The weights of the filters for each layer C are stored in an $N \times m \times 20$ array M^C, where each element $M^C_{k,j,l}$ is the weight of filter k at position j and amino acid l. Similarly, the activation thresholds for the rectification are stored in an 1-dimensional vector of length n. Putting everything together, the score for each filter of layer C at position i of an one-hot encoded input sequence S is:

$$Y_{i,k} = \max\Big(0, \sum_{j=1}^{m}\sum_{l=1}^{20} S_{i+j,l} M_{k,j,l} - b_k\Big), \tag{2.3}$$

for $i \in (1, L)$.

Global Max-Pooling After all the convolutional layers, the output is a matrix Y of dimensions $L \times N$. To loose the dependence on the protein length, and map all protein sequences into embeddings of a fixed dimensionality, we take the global maximum for every filter across an input sequence. This operation is actually fundamental in the construction of motif embeddings, as it dictates that each filter learns a single protein sequence motif, which in turn allows the visual interpretation of motif embedding components. Eventually, the output of the convolutional plus Max Pooling stage is a one-dimensional vector X of length N, i.e. for a input protein sequence S:

$$X(S) = gMP\big(ConvNet_M(S)\big) = (x_1, x_2, \ldots x_n). \tag{2.4}$$

Fully Connected Network The last piece of the core **CoMET** architecture is a fully-connected (FC) network of F layers. For all purposes discussed in this thesis, the number of FC layers within the core architecture was either one or two. Each layer, has a weight matrix $\mathbf{W}$ and an activation threshold vector $\mathbf{b}$. The output dimension of the FC network, which essentially corresponds to size the motif embedding $\mathcal{M}$ of input sequence S, is equal to the number of filters N in the convolutional stage[4]. In

[4]We experimented also with a large number for the output dimension to enforce sparsity in the embeddings (sparse coding) but the results did not look promising at the time of writing of this manuscript.

the case of a single FC layer, the motif embedding of a sequence S is computed as

$$\mathcal{M}(S) = FC_W\Big(gMP\big(ConvNet_M(S)\big)\Big) = \mathbf{f}(\sum_{i=1}^{N} \mathbf{W}_i X(S)_i + \mathbf{b}), \qquad (2.5)$$

where $\mathbf{W_i}$ is the i-th row of the weight matrix, $\mathbf{b}$ is the vector of activation thresholds and $\mathbf{f}(\circ)$ is a non-linear function applied element-wise to its argument vector. Depending on the application of the motif embeddings, $\mathbf{f}$ is either a *sigmoid* or a $ReLU$, and the rationale behind the choice is explained in each case.

2.2.2 Unsupervised and Supervised Learning Extensions

In this section, we describe two exemplary network architectures that take as input motif embeddings, i.e. the output of the final layer of a **CoMET** model. The first one, which we call **CoDER**, extracts motif embeddings from a set of protein sequences, where no other information is available, in an unsupervised way. Essentially, the complete **CoDER** architecture is a stacked convolutional auto-encoder [33], which we create by inverting the core **CoMET** architecture and connecting it to the motif embeddings layer. As you can see in Figure 2-4, the decoding "extension network", is the mirror image of the encoding network.

The second architecture, which we call **CoFAM**, allows us to learn motif embeddings from protein sequences in a supervised learning setup, when, for example, protein family information is available. Here we extended the core **CoMET** architecture with a set of fully-connected layers, leading to a final $Softmax$ layer that performs the family classification task (Figure 2-5). Practically, this architecture can be used to associate the motif embeddings with any label information available for the training dataset. The latter is fundamental in learning meaningful protein representations that can be used to describe the differences and commonalities across a set of protein sequences.

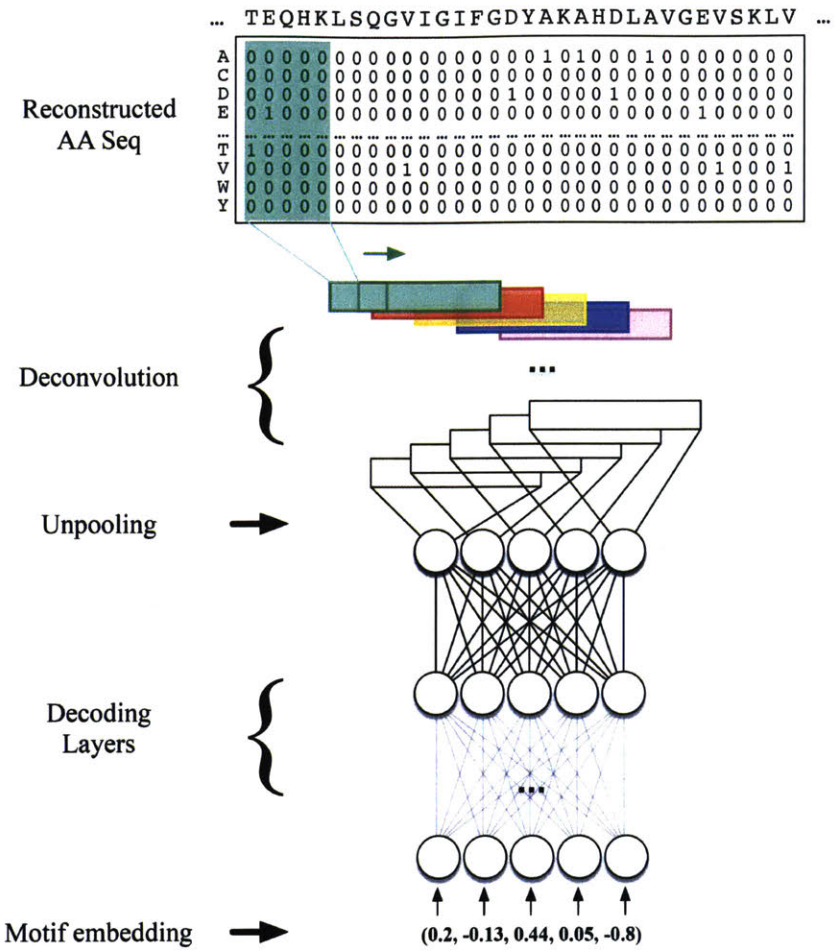

Figure 2-4: The architecture of the **CoDER** network, an extension of the core **CoMET** architecture to extact motif embeddings in an unsupervised way.

2.2.3 Training objectives and evaluation metrics

Depending on the network architecture and the training objective, we defined different metrics to be used as loss functions for the training phase. The optimization algorithm used to minimize the loss function at each step of the training procedure is in most cases *Adam* [34] with Nesterov Momentum [35], as it performed slightly better than stochastic gradient descent (SGD) in the majority of the models during hyper-parameter optimization. In many cases, we also use $L2$ weight regularization and activity regularization techniques, which add the corresponding norms to the

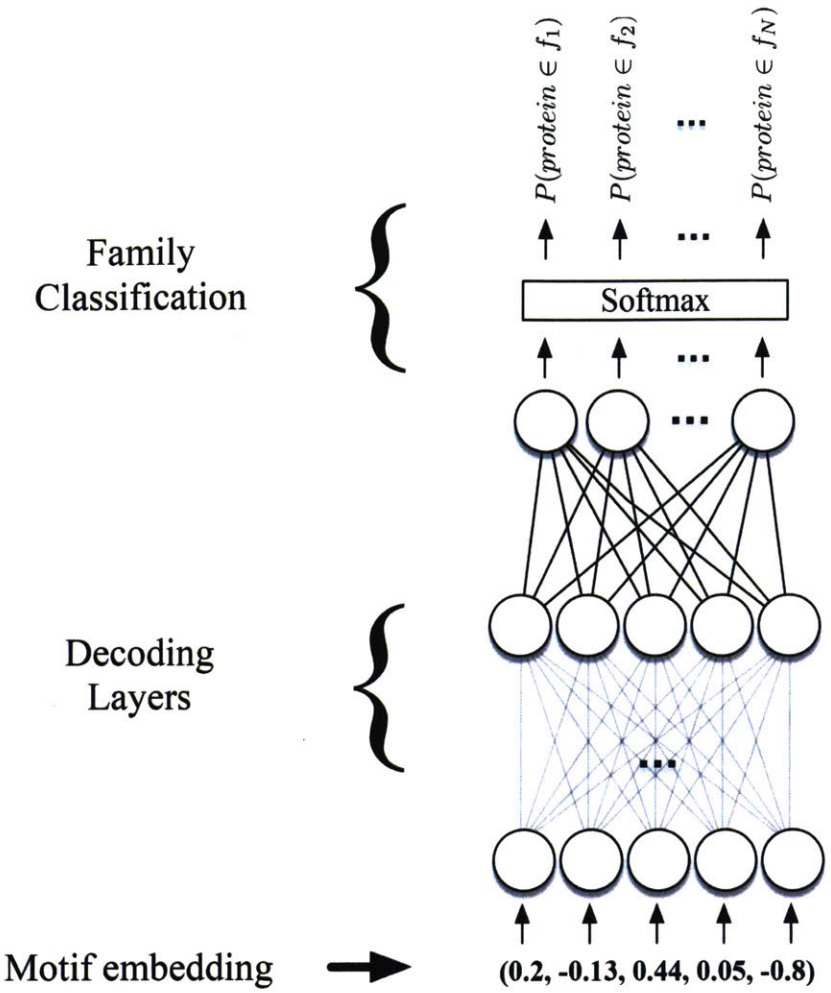

Figure 2-5: CoMET CoFAM.

training objective.

$$XLOSS = -\frac{1}{L}\sum_{i=1}^{L} H(p_i, q_i) = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j} p_{ij} \log q_{ij}, \tag{2.6}$$

where L is the length of the sequence and x cycles the 20 amino acids.

Worth mentioning is that in some cases, the categorical cross-entropy loss function was producing very large gradients that exploded the weights of the network quickly. In those cases, we found that using the *mean squared error* (MSE) results in smoother

learning and better performance overall. Illustratively,

$$MSE = \frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{20}(p_{ij} - q_{ij})^2, \tag{2.7}$$

where L is again the length of the sequence.

For the family classification network (**CoFAM**), the objective is a n-class classification, where each class corresponds to a protein family. As the output of the neural network is a typical $Softmax$ layer, here we use the well-characterized *negative log-likelihood* (NLL) as the loss function. Practically, NLL is a categorical cross-entropy of a multinomial Bernoulli distribution, where the true distribution is an one-hot encoding of the classes. Thus, if q_i the output of the i-th $Softmax$ neuron,

$$NLL = -\sum_{i=1}^{N} \mathbb{1}_{ij} \log q_i = -\log q_j, \tag{2.8}$$

where j is the index of the correct class.

Apart from the loss functions, we also define a set of four performance metrics used to evaluate both the generalization abilities of the neural network, as well as the motif embeddings per se. These metrics, are used for the evaluation of **CoMET** architectures during the hyper parameter optimization.

Sequence Reconstruction Score The first metric, which we call *Sequence Reconstruction Score* (SRS), is implemented for the case of the auto-encoder architecture (**CoDER**), and corresponds to the mean categorical accuracy, i.e.

$$SRS = \frac{1}{L}\sum_{i=1}^{L} \mathbb{1}(\arg\max_j p_{ij} = \arg\max_j q_{ij}), \tag{2.9}$$

where j cycles through the 20 amino acids. For a given sequence dataset, SRS is calculated for each sequence and then averaged to obtain the SRS of the underlying network architecture.

Family Classification Score Similar to the SRS score, the *Family Classification Score* (FCS) corresponds to the categorical accuracy for the n-class family classification. If q_i the output of the i-th $Softmax$ neuron and n the index of the correct class, then for a group of m sequences

$$FCS = \sum_{i=1}^{m} \mathbb{1}(n = \arg\max_j q_{ij}), \tag{2.10}$$

where j cycles from 1 to the total number of classes N.

Motif Information Score Given a trained **CoMET** network, we extract a set of sequence motifs in the form of *Position Weight Matrices* (PWMs). The PWMs are generated by aligning the regions across the input dataset that activate each filter of the final convolutional layer (see Appendix A.2 for a detailed description of the motif extraction method). The information content of the PWMs is used as a performance metric, as the major driver behind the motif embedding representation is the interpretability of the learned embeddings in the input sequence space. For a set of M motifs of length L, extracted from a network with M filters in the final convolutional layer, the *Motif Information Score* (MIS) is the average information content per motif, or

$$MIS = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{L} \sum_{k=1}^{20} -p_{ijk} \log p_{ijk}, \tag{2.11}$$

where p_{ijk} is the probability weight of the k-th amino acid (out of 20) in the j-th position of the i-th motif.

Homology Reconstruction Score Last but not least, an important property of the learned motif embedding representation is to preserve faithfully the distances of the input protein sequences within a known protein family. In other words, protein homologs should have neighboring embeddings and should produce a meaningful phylogenetic tree with a hierarchical clustering algorithm. To evaluate the generated phylogenetic tree, we define the $HomologyReconstructionScore$ (HRS), which

is essentially a modified version of the *cophenetic correlation coefficient*, an well-established measure of the faithfulness of hierarchical clustering algorithms. Then,

$$HRS = \frac{\sum_{i<j}(SS(i,j) - \bar{SS})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(SS(i,j) - \bar{SS})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}, \tag{2.12}$$

where,

- $SS(i,j)$ is a sequence similarity metric between input sequences i and j after alignment,
- $t(i,j)$ is the tree distance between the embeddings of sequences i and j. This distance is the height of the node at which these two points are first joined together.
- $\bar{SS}$ is the average of $SS(i,j)$ and $\bar{t}$ is the average of $t(i,j)$.

The sequence similarity metric we use for the above calculation is the following:

$$SS(i,j) = \frac{\#\text{ Similar Residues between } i \text{ and } j}{\text{Length of sequence alignment}}, \tag{2.13}$$

where residue similarity is calculated by considering groups of similar amino acids based on physicochemical properties (e.g. hydrophobicity, charge etc.).

Regularization We used weight decay ($L2$-regularization) for the weights of the convolutional layers and activity regularization for the output of the last convolutional layer. We also used the early stopping [36] regularization technique, as it is easy for the convolutional stage of **CoMET** to overfit in the training set, especially when the input sequences do not have sufficient sequence diversity.

2.2.4 Hyper-parameter optimization

Catalyst to the success of any neural network, especially in the case of deep architectures, is the optimal selection of calibration and architecture parameters such as the number of filters in a convolutional layer or the learning rate of the optimizer.

These parameters, usually called "hyper-parameters", require exquisite fine tuning and pose an obstruction to the wide-spread use of deep learning. To optimize the hyper-parameters of **CoMET**, we used state-of-the-art techniques [37] with significant success in networks trained on text and images.

CoMET has two main sets of architecture-related hyper-parameters that required optimization, namely

1. the number and length of *motifs* (filters)[5]
2. and the number of convolutional and fully connected layers.

Number and length of motifs. The size and number of motifs can vary significantly based on the characteristics of the protein sequences. In Figure 2-6 you can see the distribution of lengths for all documented protein motifs in the PROSITE database. There is a peak around 15, but there are motifs more than 200 amino acids long.

Motifs should be short enough so that the sequence reconstruction problem does not become trivial. On the

Factors that influence the number and length of filters:

- Biological interpretation of the motif. Usually if a motif implies structural conservation it is in the order of 10-15 amino acids. For sequence conservation in a particular domain, the motif length varies from 30 to even 100 amino acids in length.
- Occurrences of a motif in the same sequence.
- Sequence diversity of the input dataset. If the sequences span across multiple protein families then high number of motifs is performing better. Yet, there is a trade of
- Hierarchy of the motifs (number of layers)

[5]The words motifs and filters will be used interchangeably in the rest of this thesis. See Section 2.2.1 for details.

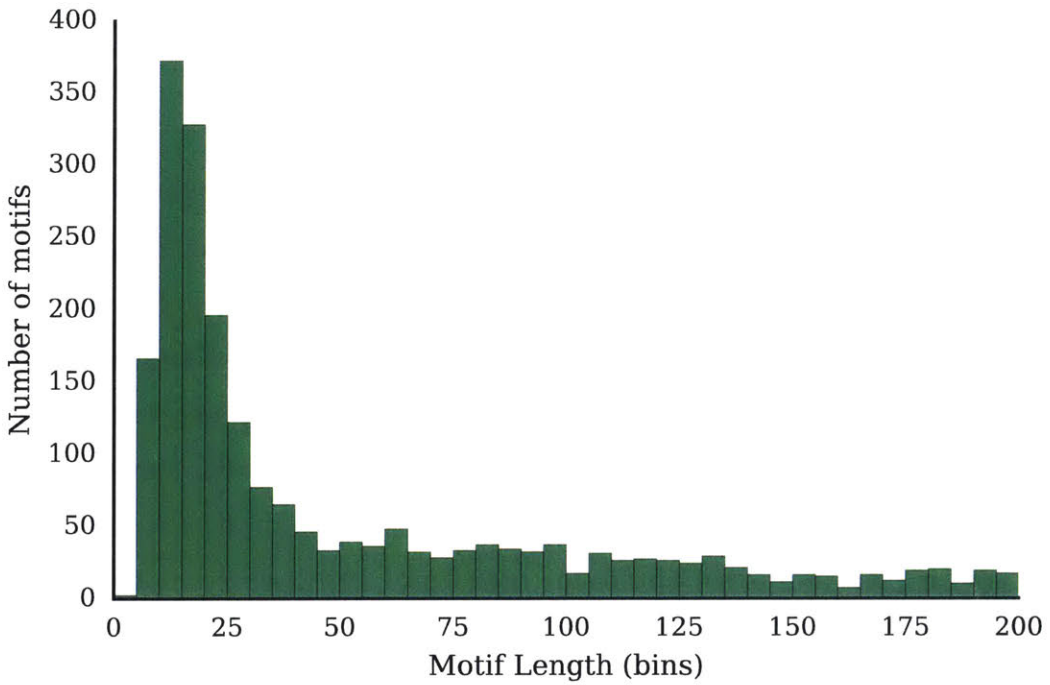

Figure 2-6: Distribution of size for protein motifs on PROSITE as of release 20.128.

In order to get an understanding of the effect of the length and number of motifs in training CoMET models, we performed a uniform sampling of the hyper-parameter space for the dataset of Cas9 proteins.

Number of Convolutional and Fully Connected Layers. The number of convolutional layers strongly depends on the size of the input dataset as it dictates the order of magnitude of the network parameters. Even with a single convolutional followed by a fully connected layer, there is a motif hierarchy emerging from the non linear combination of the first-layer motifs.

Increasing the number of fully connected layers gives the network the ability to model complex motif relationships between the input proteins, which is very useful in the case of full proteomes comprising of many families with overlapping set of motifs.

2.3 Learning motif embeddings for hierarchical clustering and family classification

After optimizing the hyper-parameters with simulated datasets, we applied **CoMET** to hierarchically decompose protein sequence datasets in motif embeddings. To start with, we investigated whether the motif embeddings could be used to recover known functional or structural protein classification. At the same time, we explored a series of hierarchical clustering techniques in order to recover the phylogenetic tree of a given protein dataset. In both cases, we compared the results between supervised and unsupervised learning approaches, which led to the conclusion that current family classification schemes fail to capture the complexity of protein evolution around a particular biological function.

To conduct the above experiments, we selected two protein super-families, namely transcription factor proteins, as they have a large phylogenetic tree with many motifs defining families and subfamilies and Cas9 proteins due to their biological importance in gene therapy and highly modular structure.

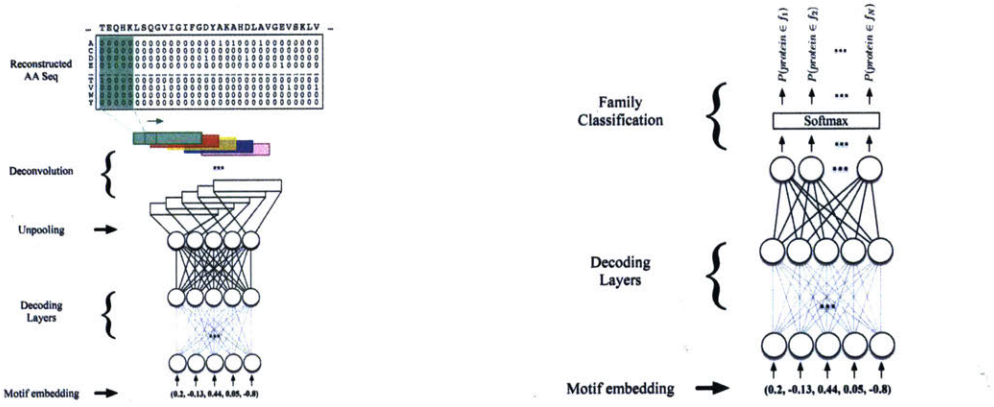

Figure 2-7: Comparison of the two network extensions of **CoMET**. The Convolutional Autoencoder (left) is trained by unsupervised learning methods to perform de-novo functional clustering of a protein dataset. For protein family classification, as supervised learning architecture (softmax regression) is used on top of the core CoMET Encoding layers.

Both architecture produce motif embeddings, that have similar properties but can

be used to answer different questions regarding the input protein dataset.

2.3.1 Transcription Factors

The first super-family of proteins we applied *CoMET* on was *transcription factors*, a large set of proteins responsible for the initiation and regulation of gene transcription. An intriguing feature of transcription factors is their modular DNA-binding domains, enable them to bind to specific sequences of DNA. Thus, we asked the the question: are all transcription factors composed by a fixed set of motifs associated with DNA-binding function? Our hypothesis was that the motif embeddings of transcription factors will capture this modularity and provide insights for their evolutionary composition and phylogeny.

We begun our analysis by training an unsupervised network (**CoDER**) on a set of 11036 transcription factors collected from Uniprot (see Appendix A.1 for the detailed query). The metric used as a training objective was the *Sequence Reconstruction Score* (SRS). To assess the performance of the network, we randomly split the dataset into training (80% of proteins) and validation (20% of proteins) sets.

The training curves for one of the best performing **CoDER** architectures on the transcription factor dataset are shown in Figure 2-8. The architecture had two convolutional layers followed by two fully-connected layers. The optimal number of filters, and thus the length of the motif embeddings, was 400 with a filter length of 50.

The defining characteristic of all **CoMET** architectures, is that they automatically learn motif detectors and consequently rules to combine them into motif embeddings. Having a pool of trained **CoMET** models with a high sequence reconstruction score, we set out to visualize the receptive fields of the convolutional network neurons, in order to identify which parts of the input sequences triggered the highest neuron activations, and thus were mostly reconstructed at the output. Intuitively, those parts would correspond to the most conserved regions among the training dataset sequences, and thus contain information about the protein families. In Figure 2-9, we show the top eight extracted motifs from the convolutional layer of the trained

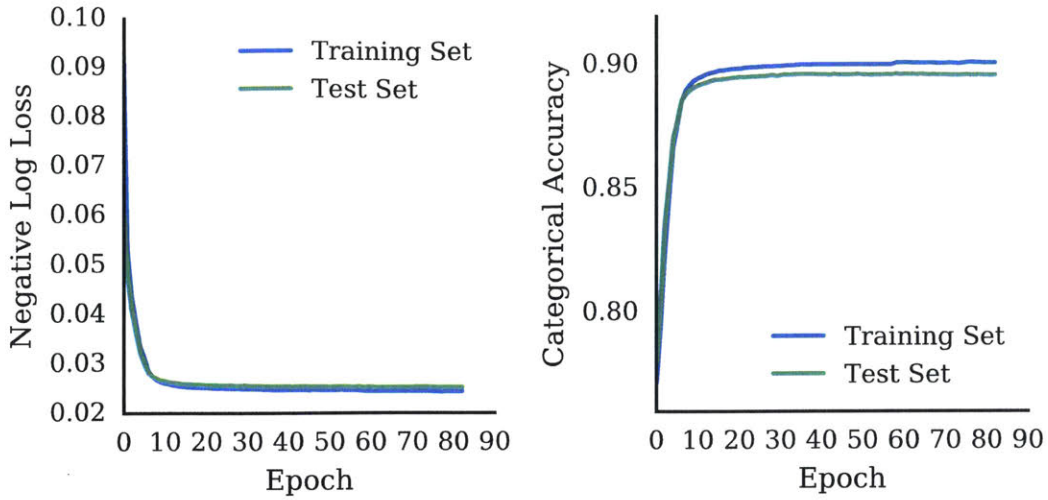

Figure 2-8: Loss and accuracy curves for training and validation sets of one of the best **CoDER** models on the transcription factor dataset. The number of epochs was dictated by early stopping. The loss function is the Mean Squared Error (MSE) and the accuracy metric is the Sequence reconstruction Score (SRS).

model. To the confirmation of our hypothesis, these motifs correspond to well-known DNA transcription factor signature profiles, such as $C2H2$ zinc fingers, $RING$-zinc fingers and homeobox domains.

To understand whether the learned motif embeddings offer a meaningful representation of the input protein sequences, we visualized the embeddings in two-dimensions using *t-SNE* [38]. t-SNE is a widely-used, stochastic dimensionality reduction algorithm that tries to minimize the Kullback-Leibler divergence between the probability distributions neighborhoods in the high dimensional and the low dimensional space. As you can see in Figure 2-10, the algorithm resulted in the formation of several clusters in the embeddings space. Coloring each data-point (protein) based on the protein family annotation, it is easy to see that the motif embeddings clusters correspond to evolutionary sequence conservation within protein families. Importantly, due to the **CoMET** architecture, we can extract a set of distinct modular motifs or conserved regions across the members of each cluster.

As a last step to our analysis of the transcription factor dataset, we further investigated the power of motif embeddings by examining whether they captured the

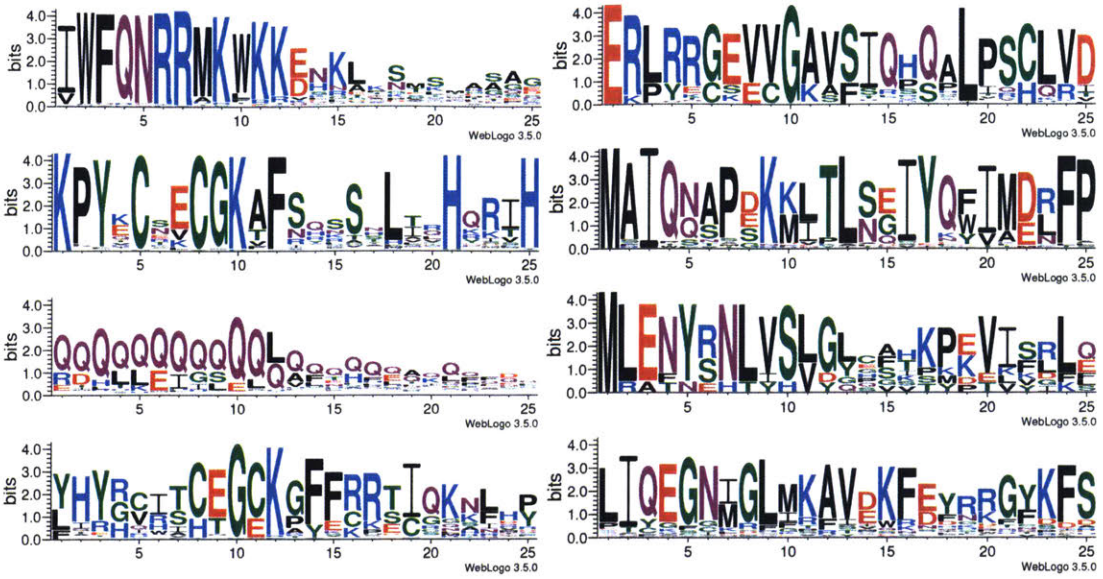

Figure 2-9: Receptive fields of the **CoDER** architecture visualized as sequence logos. Essentially, the motif embeddings can be viewed as non-linear combinations of the extracted protein sequence motifs.

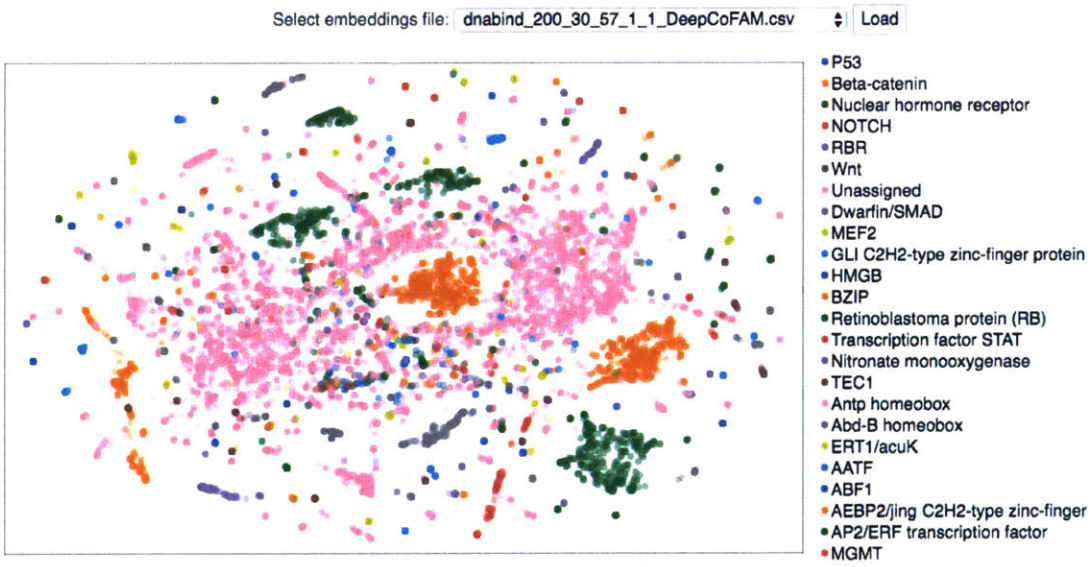

Figure 2-10: Visualization of the motif embeddings of transcription factors using t-SNE.

phylogeny of transcription factors. We started by calculating the distance between the (high-dimensional) motif embeddings for all protein pairs to form a distance matrix D. Subsequently, we used an agglomerative (hierarchical) clustering algorithm

to compute a linkage matrix, which we visualized as a dendrogram (Figure 2-11. The resulting dendrogram corresponds to a phylogenetic tree for the transcription factors, successfully grouping together proteins of the same families, as well as maintaining the evolutionary distance between different families.

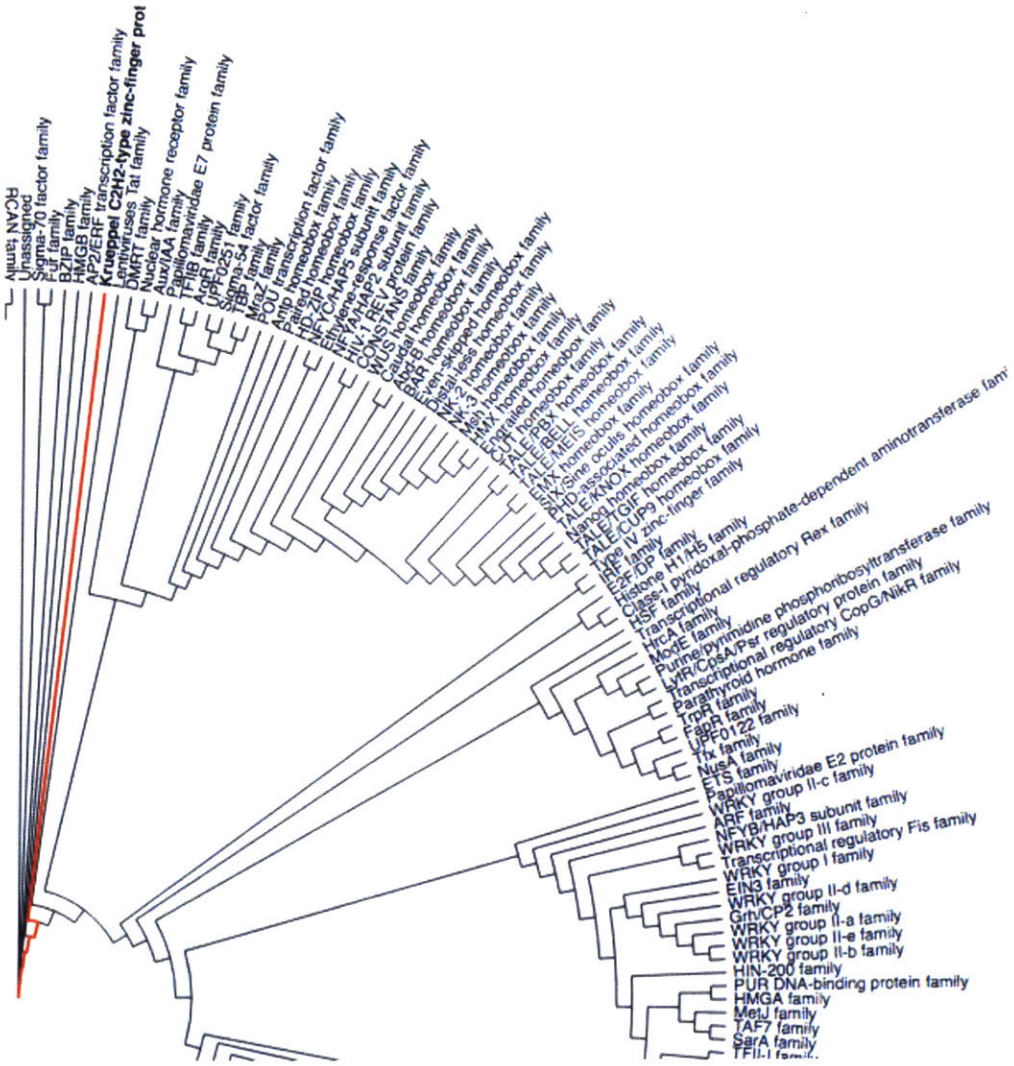

Figure 2-11: A quadrant of the phylogenetic tree of transcription factors created using hierarchical clustering on the motif embedding representations. As apparent from the figure the motif embedding captured the homology of Homeobox domain proteins.

Conclusively, in this section we trained **CoMET** on DNA transcription factor motifs and demonstrated that using the learned motif embeddings representation, we can faithfully reconstruct the intra- and inter- family relationships among the input

protein sequences.

2.3.2 CRISPR Associated Endonucleases

Staphylococcus Pyogenes Cas9 (SpCas9) [39] was the first and still most widely used *CRISPR ASsociated* (CAS) nuclease for gene editing. SpCas9 has complex structure (Figure 2-12), which is comprised of more than 1300 amino acids, has been optimized by random mutations and environmental selection to recognize a specific DNA sequence (directed by an RNA molecule) and subsequently cleave the double stranded DNA within that sequence. Due to this structural complexity, sequence homology methods have been used to identify variants with desired properties e.g. smaller protein size or higher binding affinity. The first result of the sequence homology search using *BLASTp* was Cpf1, a Cas9 analog which maintains the RNA guided nuclease activity but is much smaller in size that SpCas9. Yet, global sequence homology is not sufficient to look for CAS proteins across thousands of years of evolution, as CAS proteins are the result of recombination of several distinct submodules, each with a specific function and evolutionary history.

On the contrary, training **CoMET** on CAS proteins, we can extract the conserved motifs of those submodules and associate each CAS protein with a motif embedding, i.e. a non-linear combination of the individual motifs. Subsequently, we can use the trained network to extract the motif embeddings of all available protein sequences, and see whether they match those of the already known CAS proteins. Furthermore, we can extract the individual conserved motifs from the filters of the convolutional layer, in order to identify the submodules of the known CAS proteins and start synthesizing new variants by recombining the submodules in novel arrangements, not found in nature.

To validate our working hypothesis, we applied **CoMET** on 9286 CAS protein sequences (see Appendix A.1 for the Uniprot query), translated from the encoding regions of thousands of bacterial genomes, available through the latest NGS experiments. To start with, we first experimented with an unsupervised learning architecture in order to identify functional clusters within the CAS protein family. In Figure

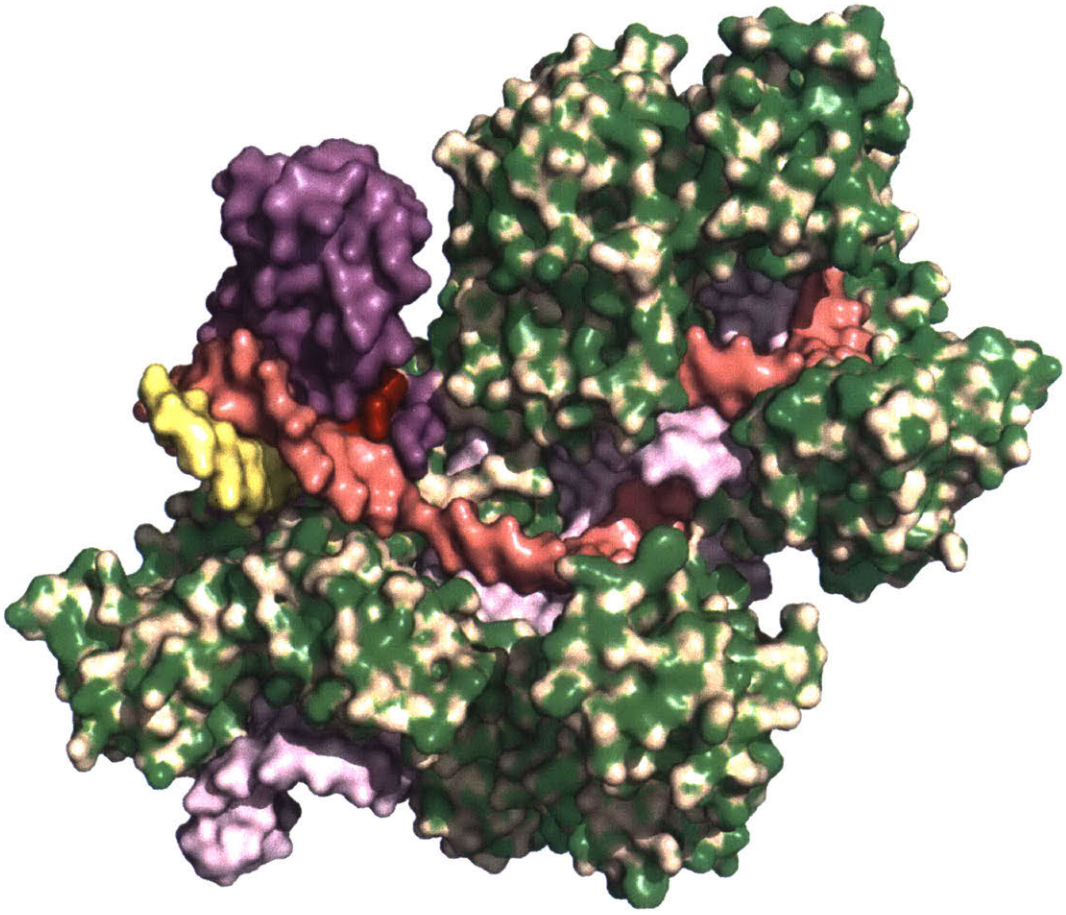

Figure 2-12: One of the solved SpCas9 CRISPR Associated protein structures ($4UN3$) visualized with a surface model using PyMOL.

2-13, you can see the loss and accuracy for a uniformly split training and validation set. Naturally, we expected to see the different Cas9 families in distinct clustered, but with CoMET embeddings we were able to create a cluster hierarchy even within a family.

Using the motif visualization technique on the receptive fields of the last convolutional layer, we extracted eight major motifs (Figure 2-14). Searching the motifs against known motifs databases (*PROSITE* and *PFAM*), we identified the HNH-nuclease motif, as well as two RuvC motifs. Most of the sequence motifs correspond

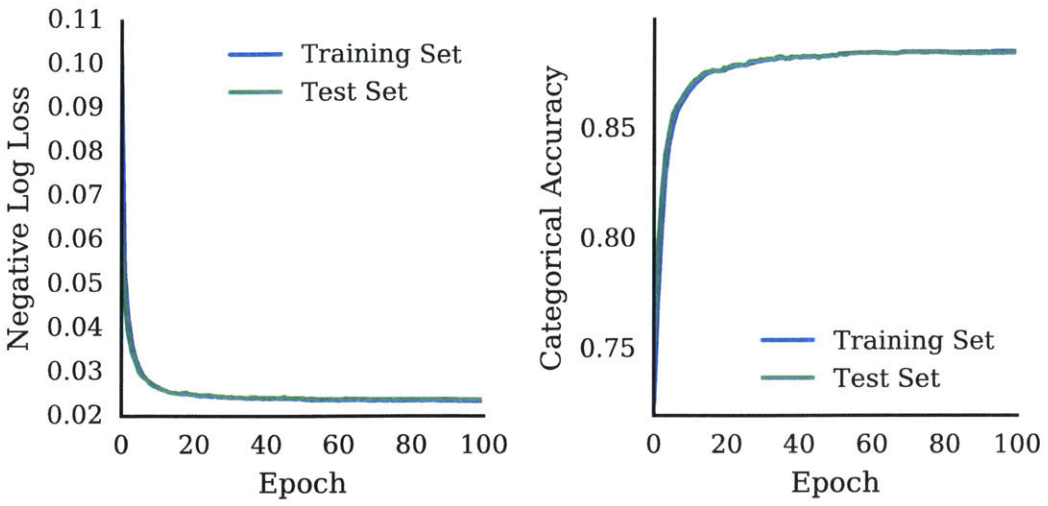

Figure 2-13: Loss and accuracy curves for training and validation sets of one of the best **CoDER** models on the CAS proteins dataset. The number of epochs was dictated by early stopping. The loss function is the Mean Squared Error (MSE) and the accuracy metric is the Sequence reconstruction Score (SRS).

to Cas9 and Cas1 proteins, as they were over-represented in the dataset.

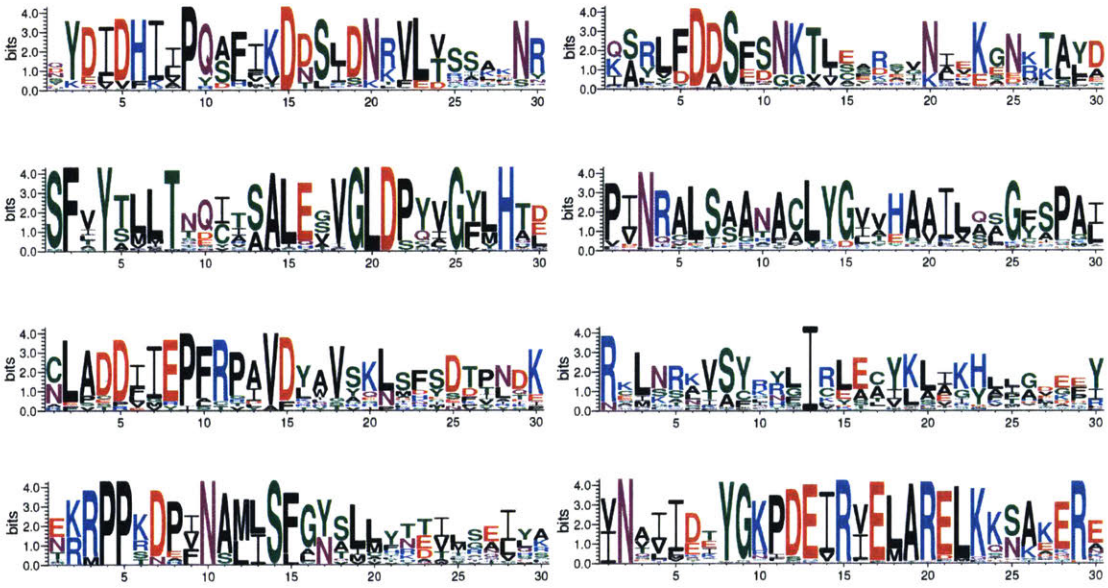

Figure 2-14: Receptive fields of a **CoDER** model trained on the CAS protein dataset visualized as sequence logos.

Subsequently, similar to the transcription factor analysis, we visualized the motif embeddings in two-dimensions using t-SNE (Figure 2-15). Looking at the visualiza-

tion, we confirm our expectations as the major protein families (Cas9, Cas1, Cas3) are in separate clusters. This is non-trivial as this was the result of an unsupervised learning method, with the objective of sequence reconstruction using a set of motifs, without any cue for protein families. We can thus conclude that **CoMET** successfully extracted the signatures in the protein sequence of each family with enough accuracy to both differentiate between families and explain the homology within.

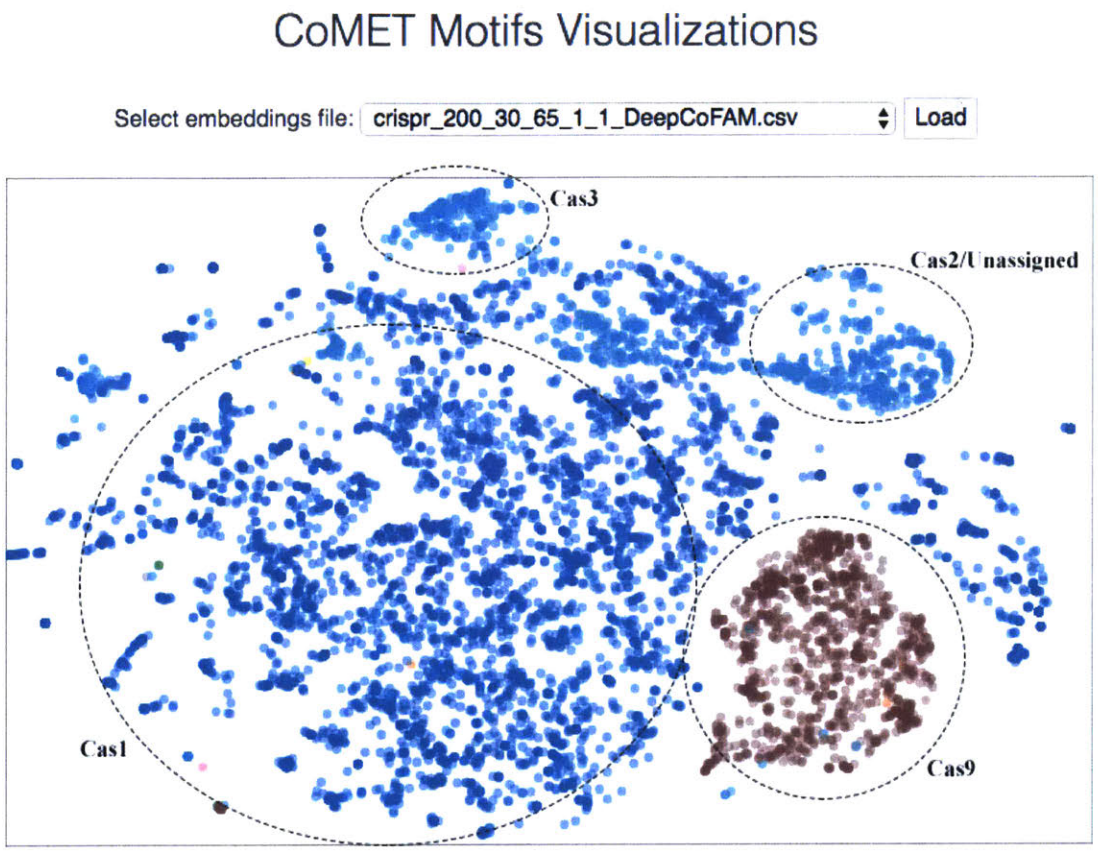

Figure 2-15: Visualization of the motif embeddings of transcription factors using t-SNE. The motif embeddings were learned using an unsupervised architecture (**CoDER**). Nevertheless, there is definite clustering of the CAS proteins in to clusters that correspond to known protein families.

Lastly, we used an agglomerative clustering algorithm to reconstruct the phylogenetic tree of the CAS proteins (Figure 2-16). Interestingly, the $Cpf1$ family was placed closer to a $CasE$ non-nuclease family, rather than a Cas9 endonuclease family. Yet, this explained by the fact that $Cpf1$ and $CasE$ families were vastly underrepresented in the compiled dataset (< 10 members each.). On the contrary, $Cas9$

and *Cas*1 families, which were also clustered correctly above, are placed in logical positions in the costructed phylogenetic tree.

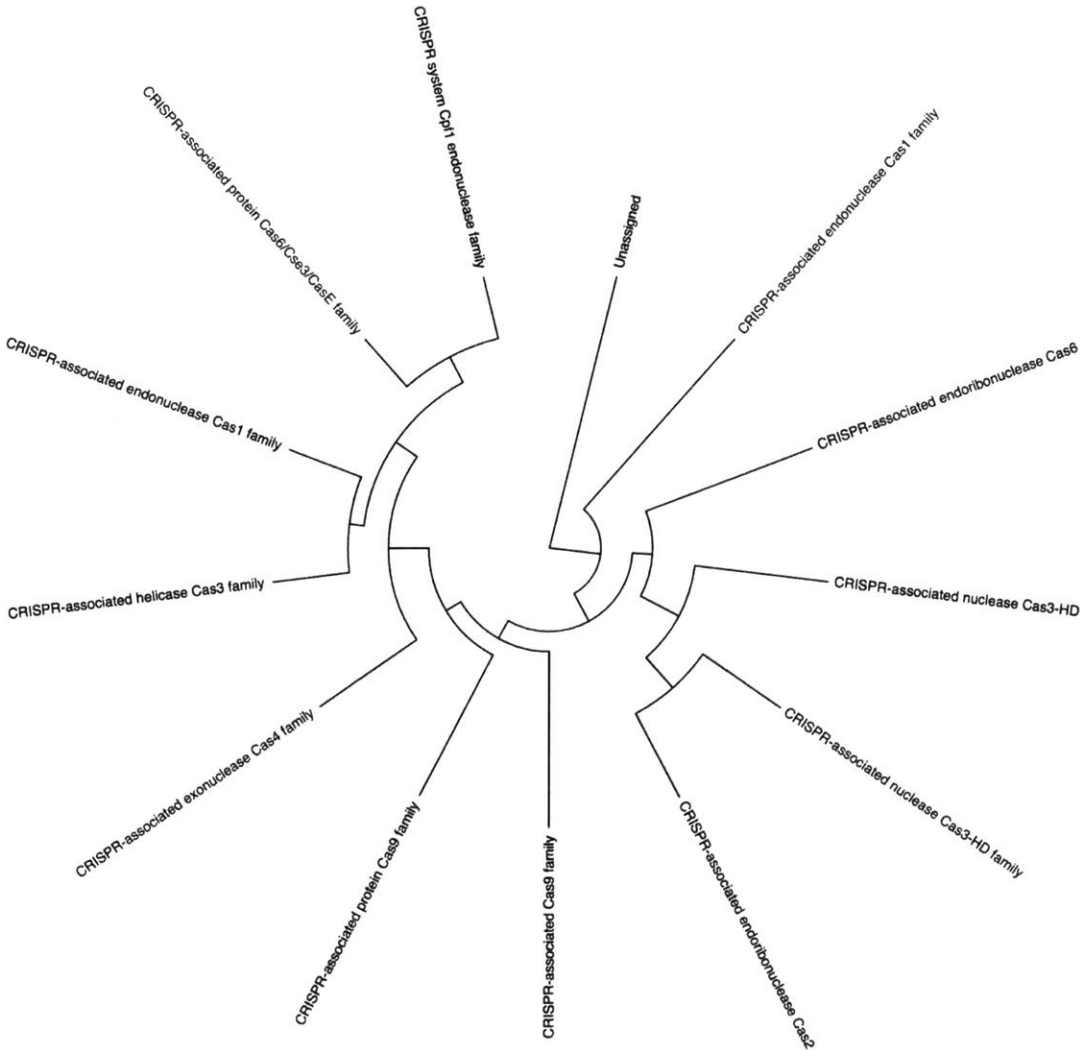

Figure 2-16: Phylogenetic tree of CRISPR-Associated proteins created using hierarchical clustering on the motif embedding representations.

2.4 Examining the phylogeny of C2H2 Zinc Fingers

So far in this chapter, we showed that **CoMET** successfully decomposes transcription factors and CAS protein super-families into motif embeddings, which can be used to cluster the proteins into functional subfamilies, with consistent intra- and inter-family relationships. In this section, we focus on extracting protein motifs from of all the convolutional layers, as well as their interconnections. We demonstrate an inherent property of deep convolutional **CoMET** models, namely the hierarchical decomposition of protein sequences into a tree of motifs of different sizes.

The dataset we assembled for this section is a collection of 8000 C2H2 Zinc Finger proteins compiled from a Bacterial One-Hybrid (B1H) binding protein selection experiment. We selected *C2H2* Zinc Fingers for their distinct zinc coordinating structural motif of two cysteines and two histidines (see Figure 2-17 for the structure and 2-18b for the sequence), which provides a reference point for the top level of the motif hierarchy.

We trained a **CoDER** architecture with three convolutional layers, and optimized the size and of filters for each layer, as well as the padding of the convolution, with respect to the sequence reconstruction score. Illustrated in Figure 2-18a is the hierarchical motif decomposition of *C2H2* Zinc Finger protein. Each layer in the tree comprises of the motifs with the highest information content, extracted from the corresponding convolutional layer. Starting from the first (input) layer with motifs of length five, the size of the conserved region that each neuron's receptive field corresponds to increases in every layer. The arrows in the figure depict the connections of the neurons between layers with the strongest weights.

Consequently, we can use the motifs' PWMs from each layer, to search for homologous proteins across protein sequence databases. As a first step, we wrote a script that transforms the PWM collection of the highest information content motifs for each layer into a file that we provide as input to *MAST*. MAST is a publicly available

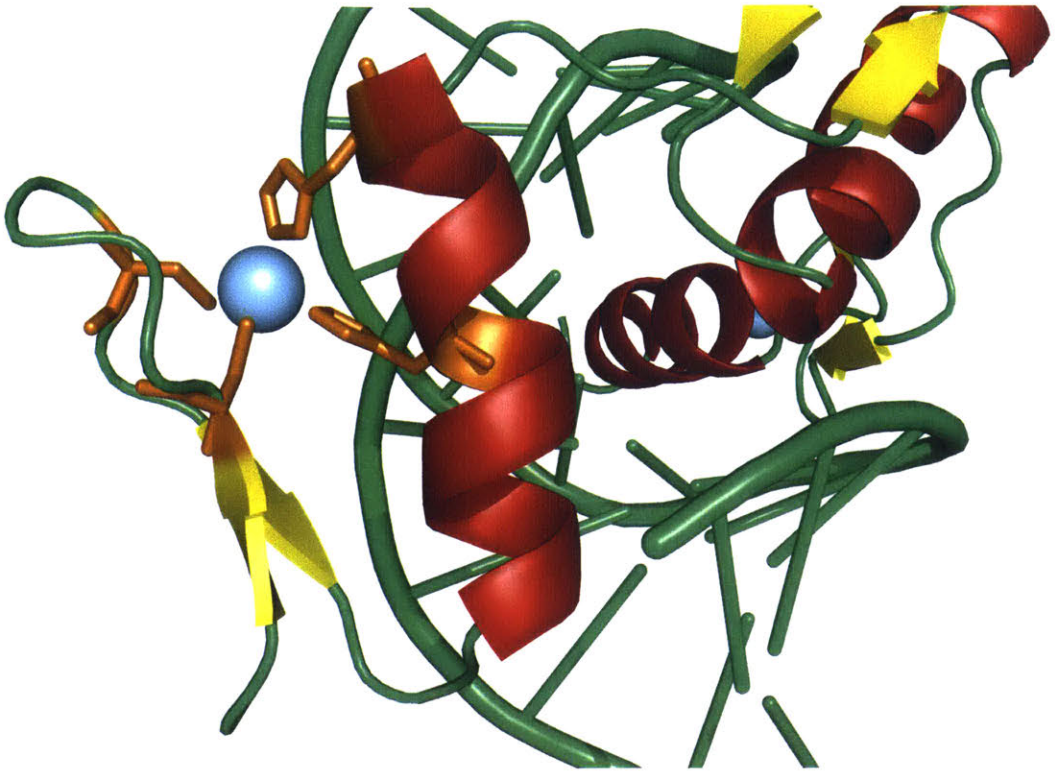

Figure 2-17: $C2H2$ Zinc Fingers were named after the zinc atom coordinating complex formed by two cysteines and two histidines. This four residue structural motif leads to the ultra conserved protein fold for this family of zinc fingers.

motif search tool, which given a formatted list of PWMs, has a set of tunable parameters to find proteins with single or multiple occurrences of one or more motifs in their sequence. Each layer of the motif decomposition tree in figure 2-18 is annotated by a keyword, which summarizes the consensus results of a MAST search with input the motifs extracted from that layer.

In Table 2.1, we have summarized the results of the MAST searches we conducted using the motifs from each layer. Analyzing the results in further detail, the search with the motifs from the first layer returned a diverse set of more than 19000 proteins, out of which the vast majority ($> 90\%$) had metal binding properties. While the specificity for zinc fingers was relatively low (45% of total number of results), the sensitivity was very high, as the results contained 99.5% of all known zinc finger protein sequences. The clear false positives in this search were a set of small proteins forming a disulphide bond between cysteines three to five bases apart, which is

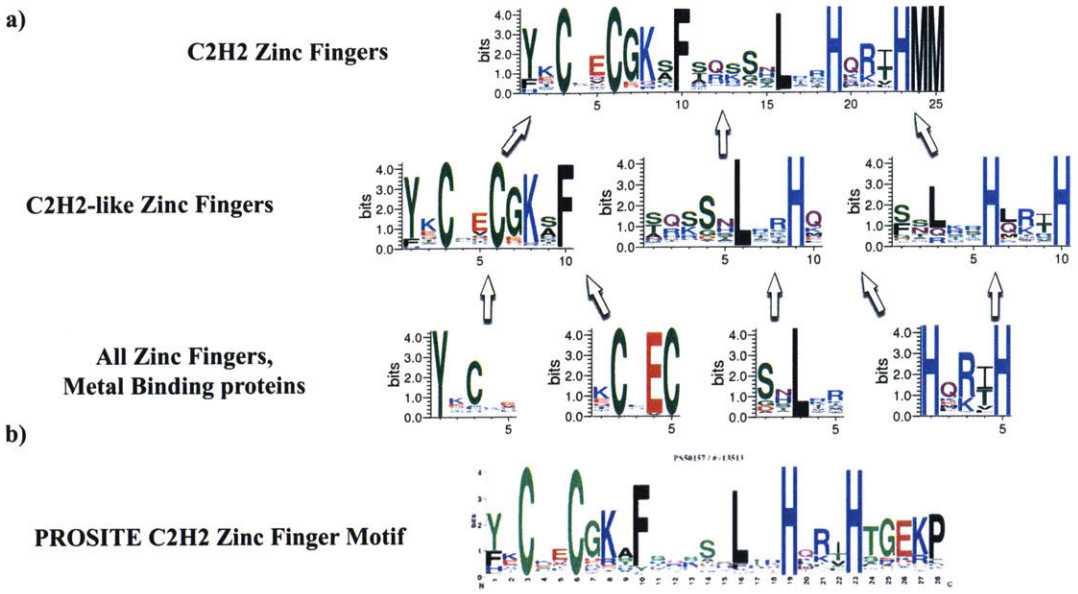

Figure 2-18: (a) Hierarchical decomposition of a set of 8000 Zinc Finger proteins into sets of motifs of different sizes. (b) The well characterized $C2H2$ Zinc finger signature profile (source: $PROSITE$)

Table 2.1: Results of a $MAST$ motif based protein search using the extracted motifs from each layer of the convolutional part of **CoMET**

Layer	Number of motifs	Results	Precision	Recall
First Layer	50	19625	45%	99.5%
Middle Layer	24	14506	68%	91%
Final Layer	10	8769	99.9%	87%

explained by looking at the two-cysteine motif of the first layer in Figure 2-18a.

The motifs extracted from the middle of the three-layer convolutional network, had a relatively higher specificity than the first layer, but crucially, the majority of false negatives were Zinc Finger proteins with fingers of type other than $C2H2$. Interestingly, this corresponds to moving backwards in time in the phylogenetic tree of the evolution of zinc fingers, since these motifs or conserved regions were part of multiple protein families. This is a first demonstration of the superiority of the hierarchical decomposition method (**CoMET**) against other motif extraction tools, since it allows an evolutionary tunability in the breadth of the search for homologous proteins.

Importantly, this tunability is based on decomposition of the conserved regions in submodules, which may correspond to different functions (e.g. metal binding). Thus, the researcher can now investigate the evolution of complex multi-domain proteins by tracing the recombination events that assembled the individual submodules.

Finally, for the final convolutional layer, the majority of the motifs extracted from the receptive fields of the neurons, were identical to the PROSITE $C2H2$ Zinc Finger signature profile (2-18b). This is not surprising, since we selected the particular dataset for its characteristic signature, to have a point of reference for the decomposition into small motifs. Hence, the search with the larger motifs had extremely high specificity (99.9% of the proteins were $C2H2$ Zinc fingers) but the sensitivity dropped down to 87% since there was a small amount of $C2H2$ Zinc Finger proteins with indels in their most conserved regions.

Conclusively, taken together, the findings presented here accent the need to revisit the notion of protein homology. Current homology search methods, hinge on global sequence identity and have maximally segregated proteins into more than $16000 families$. As a result, we have a significant number of protein sequences that do not fall under any of these families and whose function has yet to be identified or inferred. Moreover, the evolutionary events of recombination of small protein motifs are washed out by the greedy motif extraction techniques which try to maximize the size of the conserved region. Indicatively, 80% of large viral genomes have coding regions expressing protein sequences yet to be characterized. In this Chapter we have shown that using **CoMET** and the motif embedding representation of protein sequences, we can address the above limitations and provide novel protein meta-family clusters, formed by taking into account a hierarchical conserved motif phylogeny for each protein instead of a single, large conserved region.

Chapter 3

Engineering the recognition code of Type II restriction enzymes

In the previous section we introduced *motif embeddings* and examined their use in family classification and annotation of protein sequences. Yet, the identification of structural and functional motifs per se, is fundamental to protein engineering. Here, we employ *CoMET* to assist a compelling protein engineering application, namely the design of programmable restriction endonucleases.

Modern genome engineering tools still require sophisticated protein design in order to limit their off-target activity, i.e spurious cuts at genomic sites similar to the one targeted, and avoid unpredictable mutations, that might lead to cancer or cell death. On the other hand, nature-engineered DNA cutters called *Restriction Enzymes* (known also as restriction endonucleases or REases), display extraordinary sequence specificity, with little or no off-target cleavage, which renders them ideal as DNA cutting tools.

REases, though, have their own shortcoming, as we are currently limited to the protein variants found in nature with no general means for altering their recognition sites in order cut an arbitrary DNA sequence of interest. To counter this problem, we employed custom made sequence analysis tools, in conjunction with the **CoMET** framework and trained a highly accurate predictive model on the DNA recognition code of the Type II restriction enzymes. Based on the promising prediction results,

we used the trained models to generate de novo restriction enzymes and paved the way towards the computational design of a restriction enzyme that will cut a given arbitrary DNA sequence with high precision.

3.1 Background

In his 1965 seminal paper, Werner Arber established the theoretical framework of the restriction-modification system, functioning as bacterial defense against invading bacteriophage [40]. The first REases discovered while recognizing specific DNA sequences, they cut at variable distances away from their recognition sequence (Type I) and, thus were of little use in DNA manipulation. Soon after, the discovery and purification of REases that recognized and cut at specific sites (Type II REases) allowed scientists to perform precise manipulations of DNA in vitro, such as the cloning of exogenous genes and creation of efficient cloning vectors. Today, more than $4,000$ REases are known, recognizing more than 300 distinct sequences (source: `rebase.neb.com`).

Eventually, "cutting and pasting" DNA in vitro using restriction enzymes, initiated the quest for safe and scalable DNA editing in vivo to correct mutations that cause genetic diseases, a field which now goes by the name *gene therapy*. Yet, modern gene editing through site-specific cleavage moved away from the simple but limited in function restriction enzymes, towards newly discovered and engineered systems such as: Zinc Finger Nucleases (ZFNs) and Transcription Activator-like Effector Nucleases (TALENs), Meganucleases. Recently, breakthrough research on the CRISPR system, the adaptive defense system of bacteria and archaea, revealed the potential of the Cas9-crRNA complex as programmable RNA-guided DNA endonucleases and strand-specific nicking endonucleases for in vivo gene editing. In Table 3.1, we have compiled a list of the afore-mentioned gene editing tools along with their characteristic properties.

Notably, while the sequence repertoire of restriction enzymes is fairly limited with respect to the modern programmable gene editing systems, the sequence specificity

Table 3.1: Properties of major gene editing systems characterized in recent literature.

System	Protein size (kDa)	Target Site Length (bp)	Available sites	Specificity†
CRISPR-Cas9	159	23	4^{21}	++
Zinc-Finger Nucleases	60*	$18-36$	$>4^{18}$	+
TALENs	110	$13-17$	$>4^{14}$	++
Meganucleases	$20-50$	$18-22$	4^6	+++
Type II REases	$20-100$	$6-18$	10^3	++++

References: [41, 42, 43, 44, 45].

† Specificity was deduced by relative comparison between off-target rates within references.

*Data for a six finger nuclease.

restriction endonucleases remains unmatched. Thus, we set out to investigate whether we can use *motif embeddings*, introduced in this thesis (see Chapter 2), to engineer the sequence specificity of REases in a programmable fashion. Previously, Type IIS restriction enzymes have been rationally engineered to have altered specificities and a first attempt to learn their recognition code has been demonstrated [46]. On the other hand, attempts to alter the sequence specificities of Type IIP REases have been largely unsuccessful, presumably because the sequence specificity determinant is structurally integrated with the active sites of Type IIP REases. In the following sections of this chapter, we analyze the available information for all Type II REases and apply **CoMET** to represent them in the motif embeddings space, extract conserved protein motifs for each recognition site and ultimately propose a programmable way to alter their specificities.

3.2 Analyzing the Type II restriction enzymes superfamily

Traditionally, restriction enzymes are classified into four types based on subunit composition, cleavage position, sequence specificity and co-factor requirements. However, their amino acid sequences are extremely diverse, even between restriction enzymes of the same type, resulting in further categorization to several subtypes [47]. In Ta-

Table 3.2: Different types* of restriction enzymes (endonucleases).

CLASSIFICATION OF RESTRICTION ENDONUCLEASES	
Type I Cut DNA at random, far (~1KB) from their recognition sequence.	**Type IIP** Cut within a symmetric (i.e. reverse palindromic) recognition sequence.
Type II Cut DNA at defined positions, close to the recognition sequence. They are further classified to five subtypes, P, S, G, E and F.	**Type IIG** Cut outside of their recognition sequence which can be non contiguous. They have restriction and methylation subunits within a single chain.
Type IIE and Type IIF Interact with two copies of the recognition sequence and cut either only one of them (E) or both (F).	**Type III** Require two separate recognition sites in opposite orientations. Also have restriction and methylation subunits.
Type IIS Cut at a fixed position outside an asymmetric recognition sequence.	**Type IV** Cut modified (e.g. methylated DNA)

*Interestingly, Cas9-gRNA complexes from CRISPR systems can be classified as **Type V** restriction enzymes.

ble 3.2, we have summarized the defining characteristics for each of the types and subtypes of restriction enzymes. As discussed in the introduction, we focused our attention to the Type IIp and Type IIs (hereafter called "Type II") categories of restriction enzymes, mainly due to their well-defined DNA cutting positions (offering both blunt ends and overhangs) and supreme sequence specificity.

3.2.1 Sequence phylogeny

Instead of forming a single protein family, Type II enzymes are a collection of unrelated proteins of many different evolutionary backgrounds, which is apparent from their phylogenetic tree (Figure 3-1. Type II enzymes frequently differ so completely in amino acid sequence from one another, even from every other known protein, that they exemplify the class of rapidly evolving proteins, often indicative of involvement in host-parasite interactions. The first step in our research strategy is the analysis of the existing pool of endonucleases naturally found in bacteria, in order to identify whether there is an evolutionary path for the specificity of the recognition sites, i.e.

the domain of the enzyme which is responsible for the specific binding to a DNA sequence.

Type IIp REases (e.g. EcoRI), the first discovered of type II, tend to be small, globular proteins, in the 200-350 amino acid range. They cleave DNA within their recognition sequence, and often leave four base overhangs in the $3'$ side of the DNA molecule. Most type IIp REases recognize symmetric[1] DNA sequences, because they bind to DNA as homo-dimers, but a few recognize asymmetric DNA sequences, because they bind as hetero-dimers. Lastly, some enzymes recognize continuous sequences (e.g., MunI : $CAATTG$) in which the two half-sites of the recognition sequence are adjacent, while others recognize discontinuous sequences (e.g., BglI: $GCCNNNNNGGC$) in which the half-sites are separated by up to 9 nucleotides.

The next most common Type II enzymes, classified as "Type IIS", are those that cleave outside of their recognition sequence to one side. These enzymes are quite larger than others, 400-650 amino acids in length, and they recognize sequences that are continuous and asymmetric. They comprise two distinct domains, one for DNA binding, the other for DNA cutting. While they bind to DNA as monomers, they cleave DNA cooperatively, through dimerization of the cleavage domains of adjacent enzyme molecules. Thus, some Type IIS enzymes are much more active on DNA molecules that contain multiple recognition sites.

In 2009, Morgan et al. [46] identified a set of amino-acid positions within these enzymes that determine position specific DNA base recognition at three positions within their recognition sequences, by correlating between their aligned amino-acid residues and aligned recognition sequences. These findings suggest that, with subsequent analysis, we can identify an endonuclease which was re-programmed through evolutionary changes in the protein sequence to accommodate for single nucleotide changes in the recognition site. In order to pinpoint the amino-acids involved in the DNA recognition process, we further proceed with structural modeling of the recognition site in a endonuclease-DNA molecule complex.

[1]Symmetric here means that the $5' \rightarrow 3'$ DNA sequence on the forward strand is the same as the $5' \rightarrow 3'$. As a result, half of the recognition site has to be the reverse palindrome of the other, e.g. $GAATTC$.

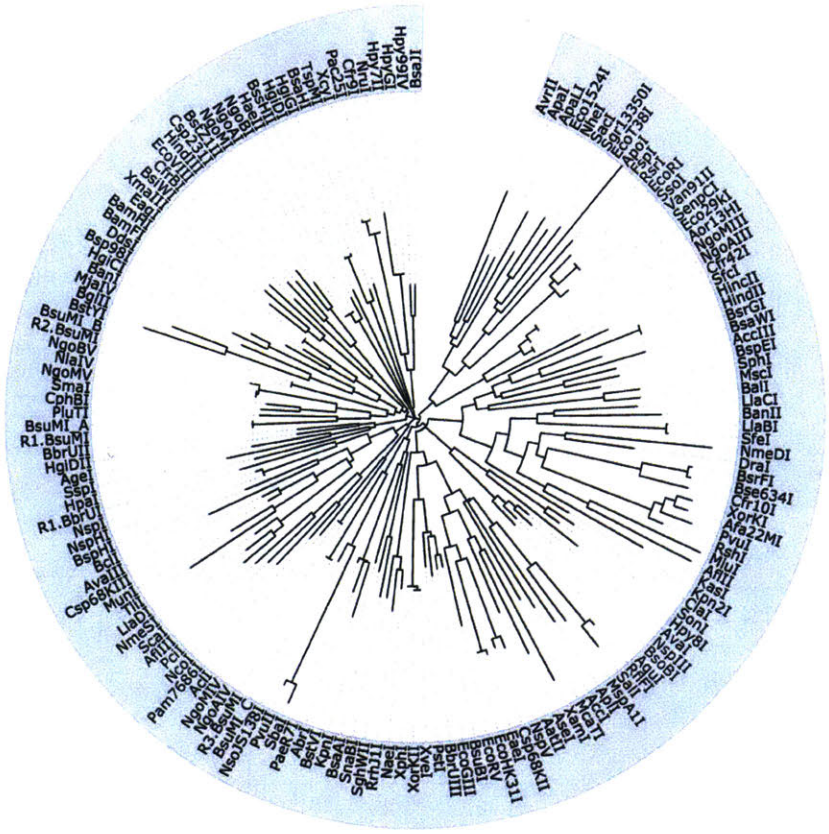

Figure 3-1: Phylogenetic tree of Type II restriction endonucleases across different bacteria species.

3.2.2 Structural inspection of the DNA binding interface

While there is a number of endonuclease structures are already available from the Protein Data Bank (PDB, `rcsb.org`), mostly the results of X-ray crystallography, there is no consistency in the structure of the different domains. With the exception of homo-dimers like EcoRI, it is generally hard to do homologous modeling of a novel endonuclease based on an already solved structure.

3.2.3 Evolution of the DNA Recognition Code

Aligning the recognition site of the protein for endonucleases that recognize neighboring sequences in the graph i.e. differing by a single nucleotide, we believe will give rise to a group of conserved amino-acids and a few select variable which made

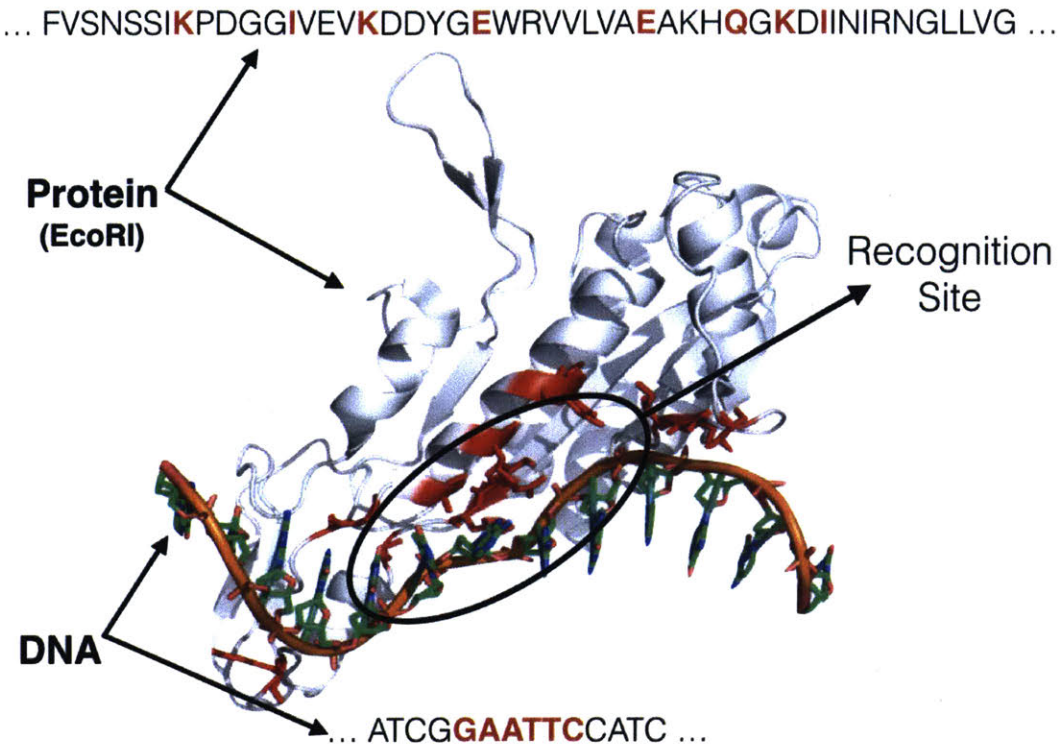

Figure 3-2: Exemplary Type II restriction enzyme EcoRI. Residues in contact (hydrogen bonding) with the recognition sequence are highlighted in red.

possible the sequence differentiation (see Figure 3-4). Subsequently, we will augment the graph with the retrieved information about the amino-acid composition in each recognition site and use information theoretic and network analysis tools to generate a set of amino-acid changes corresponding to a set of single nucleotide changes in the recognition sequence.

Initial results of generated networks are encouraging and show that a number of endonucleases recognize very similar sequences, differing by one or a small number of nucleotide bases (Figure 3-5).

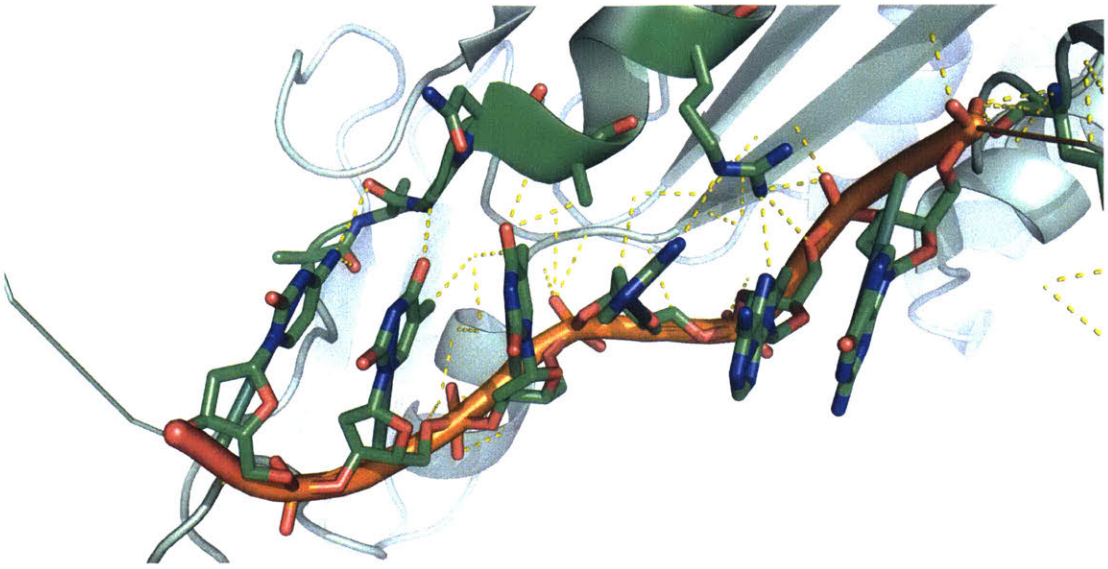

Figure 3-3: Close-up on EcoRI recognition contacts. With yellow dotter lines you can see the hydrogen bonding between EcoRI and the recognition sequence.

3.3 Inferring Type II REase DNA binding preferences from motif embeddings

We set out to solve the recognition code of Type II restriction enzymes, using motif embeddings, a hierarchical decomposition of proteins into motif combinations (Chapter 2). We compiled a large protein sequence dataset from publicly available online databases (see Appendix A.1) and designed a discriminative neural network on top of the core architecture of **CoMET**.

In total, 3595 distinct protein sequences were culled (based on recognition site availability) from the compiled dataset, out of which we further selected 2876 to train and held out the remaining 719 as an independent validation test set. To split the dataset assuring independence, we first grouped the protein sequences by recognition site, and subsequently selected at random 10% of the recognition sites for validation. The final validation test set comprised of all the sequences of the restriction enzymes that recognized the selected sites.

After an initial hyper-parameter optimization, we converged on an set of hyper-parameters which gave consistently validation accuracy above 80% (Figure 3-7).

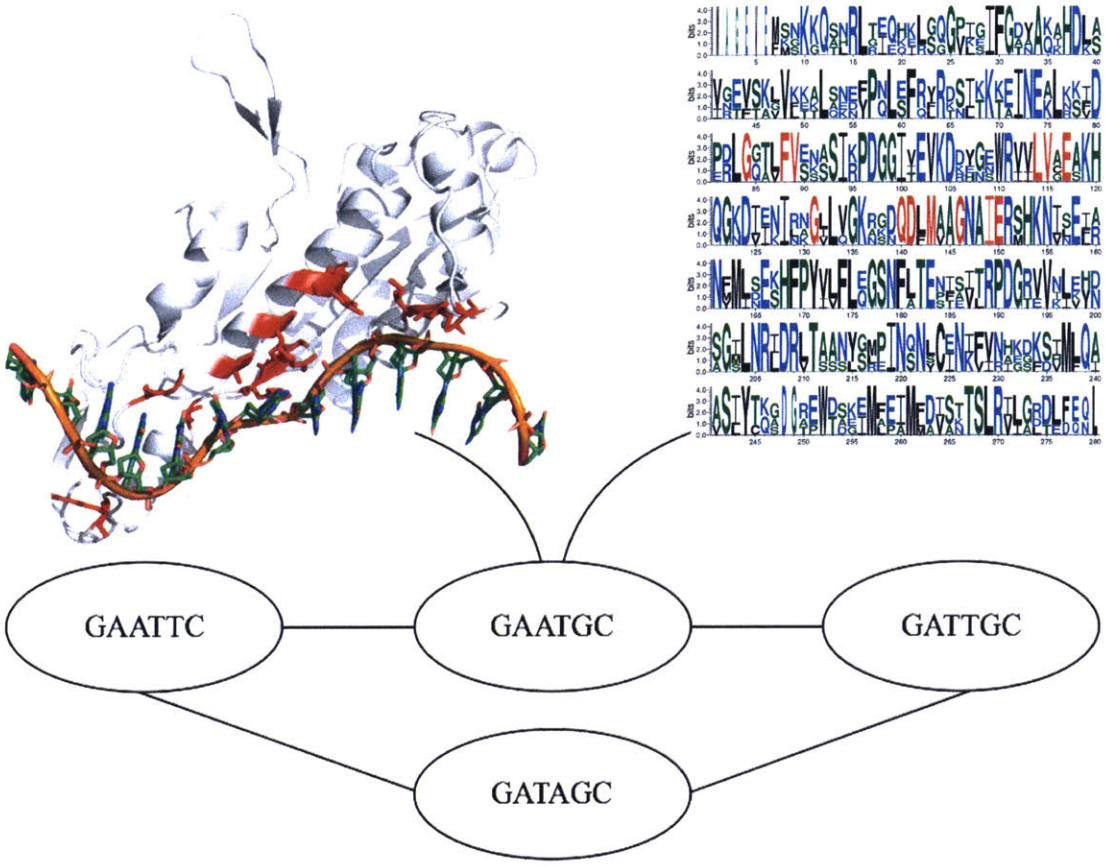

Figure 3-4: For each recognition site, we have available a set of homologs, out of which we generate the sequence logo (on the right side). If a structural model is available for any of the homologs of a particular recognition site, then we can extract the contact residues and annotate them both in the structure and the sequence logo (residues with red color).

The first observation we make, is that the discrimination accuracy differs for each position of the recognition site (Figure 3-8). This can be explained by looking at the distribution of each nucleotide (A,C,G,T) at a specific position in the recognition site within the training and validation sets. Indeed, for the worst performing base at position three, the validation set had many more sequences with a "T" at that position than the training set.

Subsequently, we visualized the experimentally determined and predicted recognition sites for a diverse set of restriction enzymes present only in the validation set, using the sequence logos of the respective PWMs. As you can see in Figure 3-9, the predictions are very close to the correct sequence.

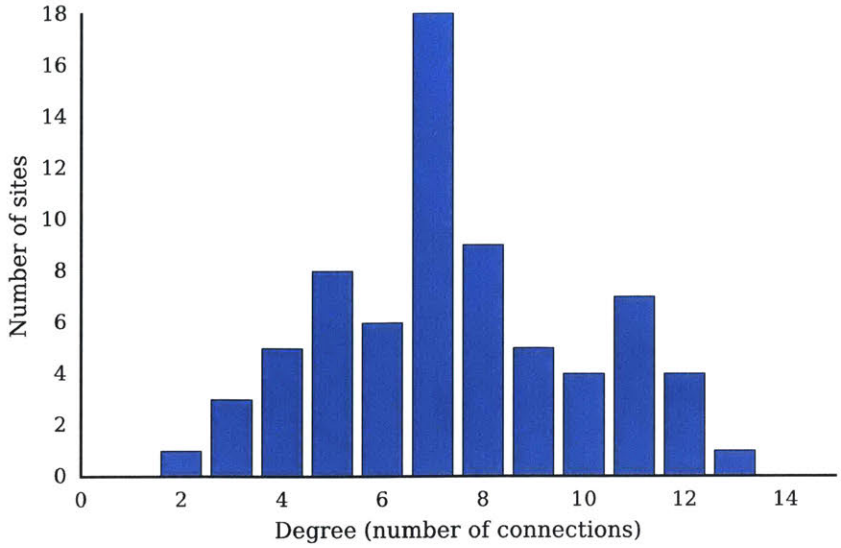

Figure 3-5: Connectivity plot for Type II restriction enzymes. Given a recognition site sequence, its degree is the number of single point mutated sequences that are still a Type II recognition site. We observe a high degree of connectivity, associated with the large diversity of restriction enzymes.

3.4 Learning to design Type II restriction enzymes with novel specificities

As a first step to gain intuition for the design of Type II restriction enzymes we visualized the receptive fields of the convolutional layers of the core **CoMET** architecture. By tracing the connections of the activated filters for a given input protein sequence all the way to the output, we start to associate nucleotides from the recognition site with motifs in the amino acids space. Thus, *CoBind* provides a set of known amino acid - nucleotide interactions which can be used as template for the engineering of restriction enzymes with novel specificities by altering the amino acids at the identified positions to those correlated with recognition of a desired new base.

Subsequently, we set out to perform a fully automated redesign of a template restriction enzyme to a target one with a different recognition sequence. Central to our method, is the ability to back-propagate the output through gradient ascent to essentially alter the input sequence towards one that would produce the given output.

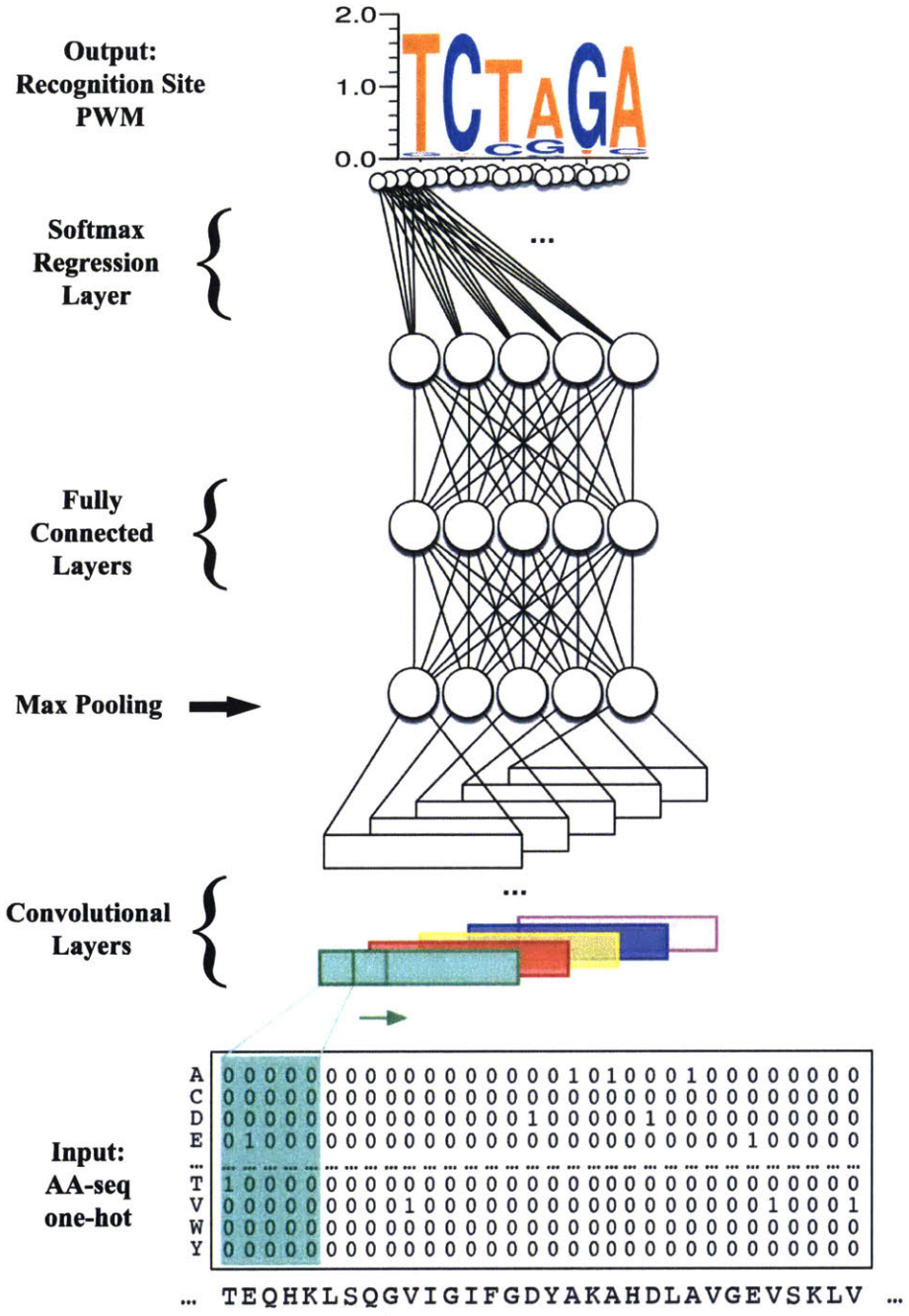

Figure 3-6: CoBind: An Deep Convolutional Neural Network architecture for the prediction of the DNA binding protein specificities based on hierarchical motif embeddings (see Chapter 2.

The developed system works by allowing the continuous evolution of a template

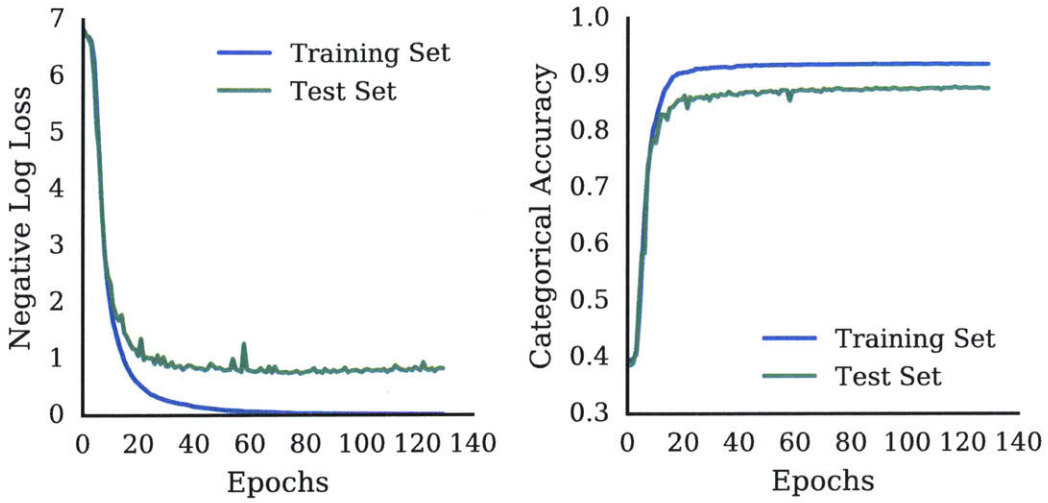

Figure 3-7: Training and validation curves for the top performing *CoBind* network trained on Type II endonuclease dataset.

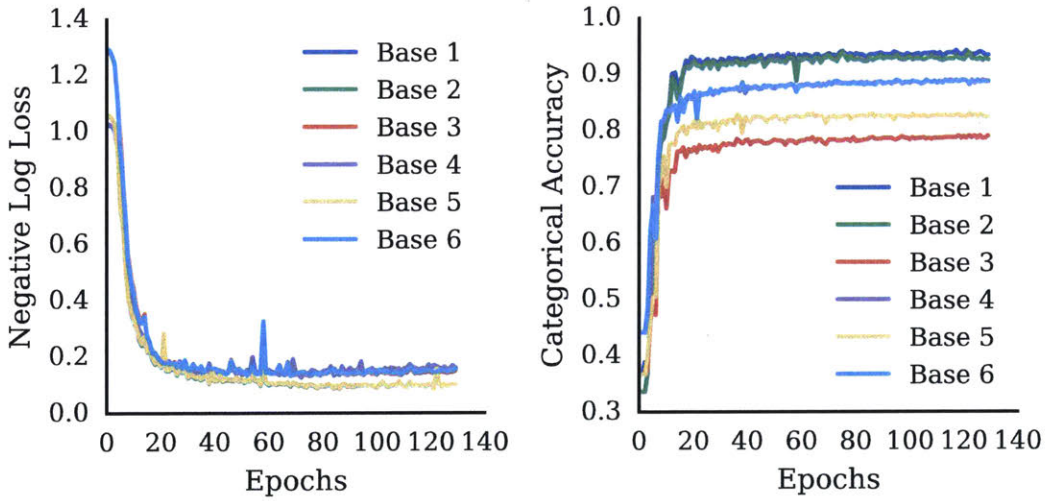

Figure 3-8: Training and test dataset performance split by base of the recognition sequence. The variation in accuracy score can be explained by the statistics of the training set recognition sites.

protein into a target protein, by a sequence of mutations in the linear amino acid sequence space. The result of the above method is shown in Figure 3-10.

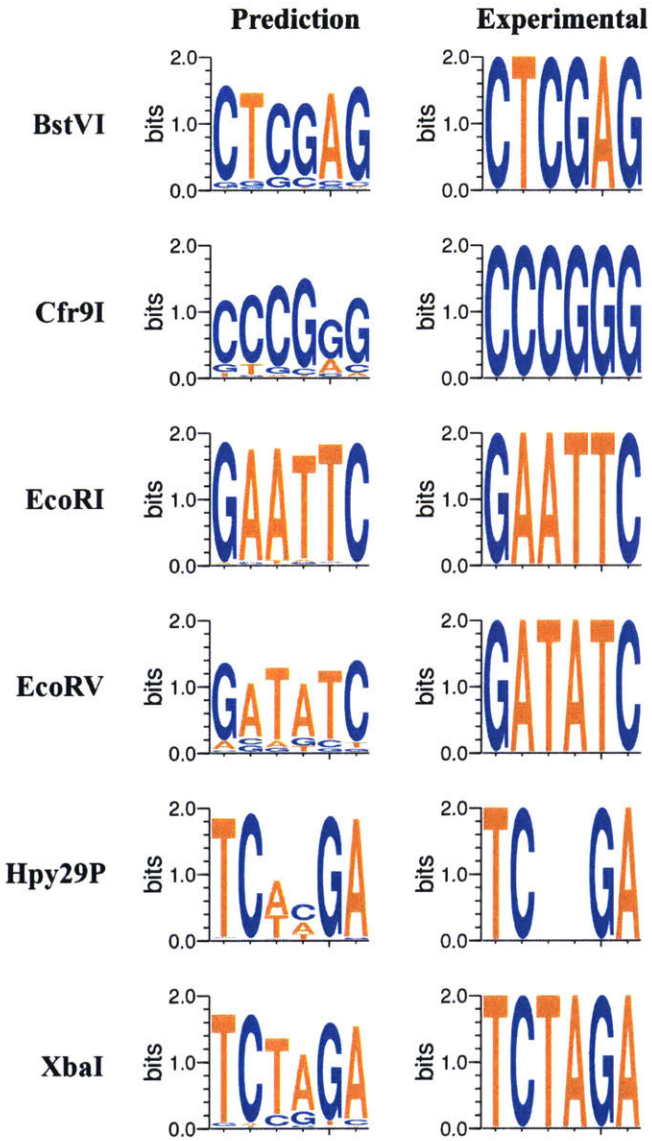

Figure 3-9: Sample outputs from CoBind. The outputs are from six held out sequences (not included in the training set).

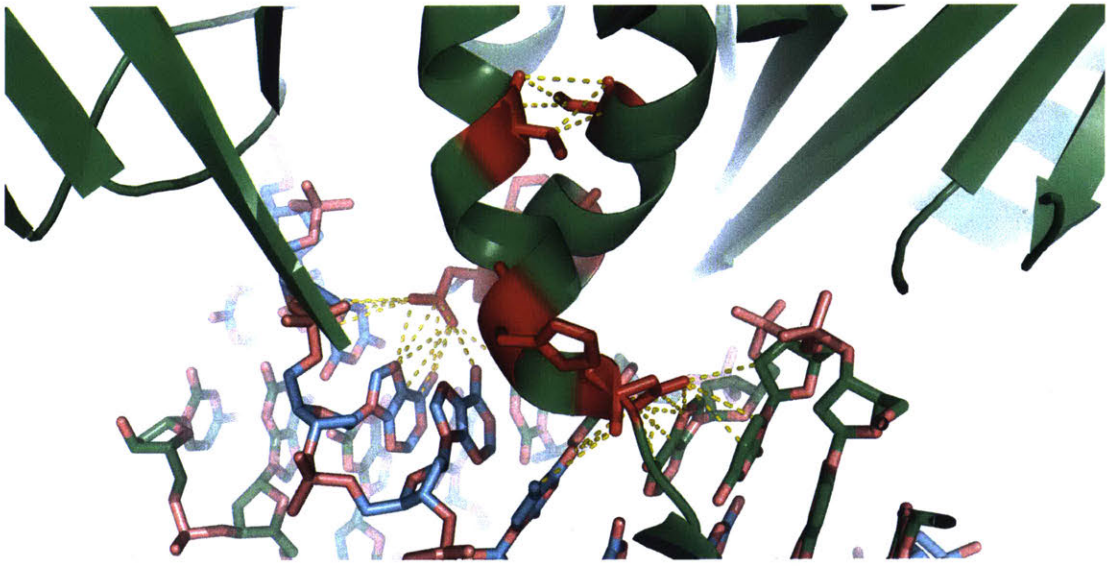

Figure 3-10: First result of generative modelling using a trained CoBind network to walk the recognition site of MunI from **CAATTG** to **GAATTC**. The residues edited are colored red. It is promising that the edited residues initiate hydrogen bonds with the recognition site, as well as assure proper dimerization (middle).

Chapter 4

Future Work

In this thesis, we displayed the power of "Deep Learning" on hierarchical motif extraction from large protein datasets. Depending on the objective function used during training, different sets of motifs were produced, each elucidating parts of the proteins that are of importance for the particular objective. Moreover, we demonstrated in Chapter 3, that it is possible to directly engineer a protein sequence towards a specific function, by reversing the extended **CoMET** architecture to perform gradient ascent on the input. While the first results look promising, we plan to further explore this technique to engineer proteins of different families.

To be confident that the engineered proteins are structurally sane and increase their probability to fold without problems, we are developing a integrated protein design framework combining **CoMET** with the well-known protein modelling software *Rosetta*. The framework allows for the simultaneous sampling of the sequence and structural space of amino acid mutations, which, in most protein enginering tasks, is extremely high for experimental screening strategies that require protein synthesis. If, for example, 10 residues within a DNA binding protein recognition site are identified as potential contacts with DNA molecule, the number of protein variants that can be generated by mutating the residues to all amino acids is 20^10.

Of course, the crucial step that follows the computational search, is the experimental validation of the engineered proteins. Throughout the work of this thesis, we took advantage of the latest advances in Next Generation sequencing to com-

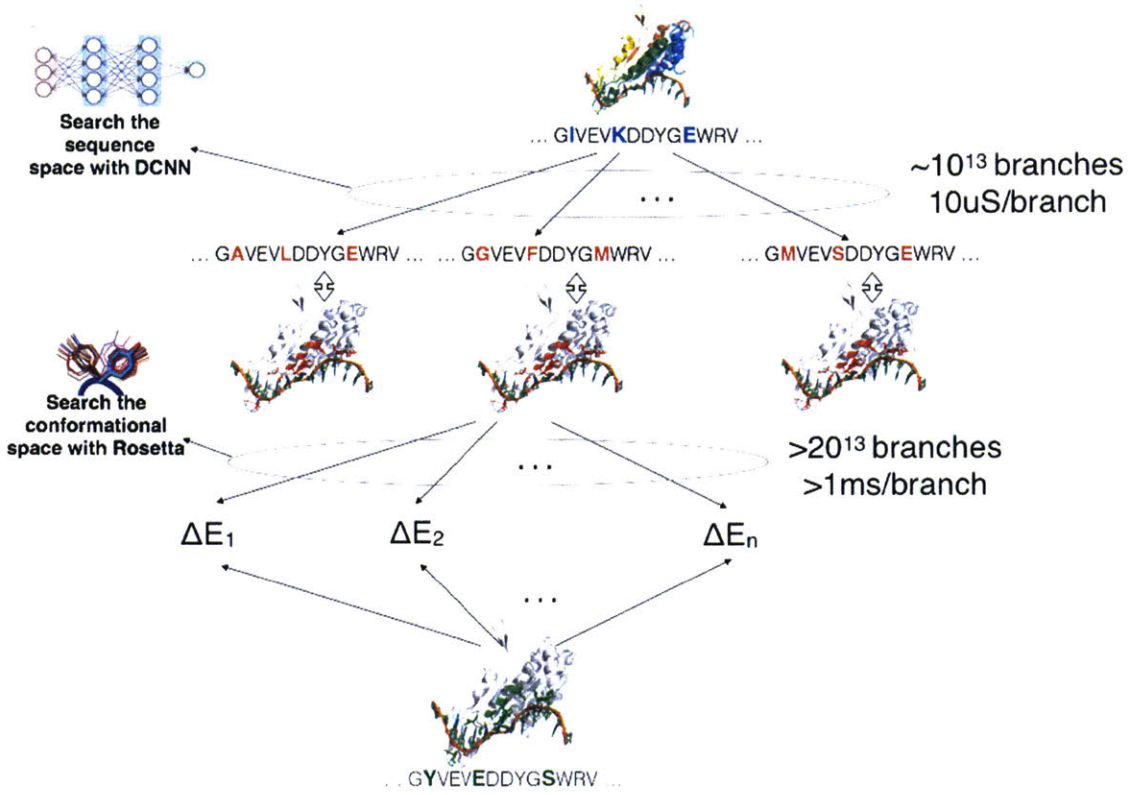

Figure 4-1: Searching for protein variants with a desired DNA binding specificity through structural modelling (*Rosetta*) and convolutional neural networks (**CoMET**).

pile large protein sequence datasets and, in Chapter 2, implemented state of the art machine learning algorithms to process them, which eventually led to the design of restriction enzyme protein variants with desired properties. Consequently, to catch up with the large pools of computationally generated protein variants, a new class of high-throughput, fast-turnaround molecular screening experiments has to be designed. Our approach (see Figure 4-2), is to synthesize a library of novel endonuclease genes with their recognition site engineered to favorably interact with a particular DNA sequence, taking into account the previously identified changes. Then, a high-throughput molecular screening experiment (negative or positive selection) will be carried out to characterize the designer proteins cleavage ability and off-target activity. Finally, the results will be used to evaluate the algorithms performance both qualitatively and quantitatively, by using the experimental sequence-specific binding

affinities to inform the energy-based structural modeling part of our framework.

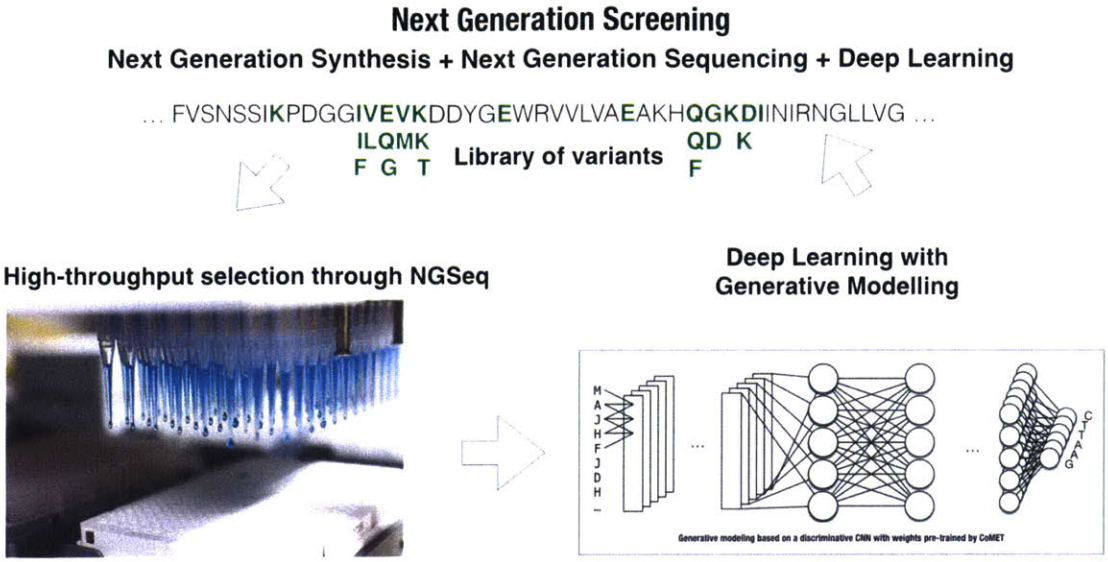

Figure 4-2: Next Generation Screening: Advances in sequencing (*NGS*) and data analytics (deep learning) lead to the need for a new class of high-throughput, fast-turnaround molecular screening experinents.

Appendix A

Methods

A.1 Data Acquisition and Pre-Processing

Protein sequence datasets were collected from *Uniprot*2016$_0$7(`http://www.uniprot.org/uniprot/`) by searching the database with custom queries. After download, most datasets were subsequently curated, as the search results contained various spurious entries (e.g. uncharacterized proteins, protein fragments etc.). The datasets used and cited in this thesis were the following:

Zinc Finger Proteins	
A list of all experimentally validated proteins which have a Zinc Finger domain.	
Query:	`zinc finger length:[50 TO *] AND reviewed:yes`
Results:	13887

Homeobox Proteins	
A list of all experimentally validated proteins that have a Homeobox domain.	
Query:	`homeobox length:[50 TO *] AND reviewed:yes`
Results:	1786

Cas9 Proteins	
A list of protein homologs of Cas9.	
Query:	`cas9 length:[50 TO *]`
Results:	9286

Complementary to the above where the following datasets collected from recent experimental literature and other sources.

Type II Restriction Enzymes	
A list of all Type II Restriction Enzymes and homologs.	
Query:	`go:"Type II site-specific deoxyribonuclease activity [0009036]" length:[50 TO *]`
Results:	4028

DNA Binding Proteins	
Polymerases, TFs, endonucleases and more.	
Query:	`annotation:(type:dna_bind) length:[50 TO *] AND reviewed:yes`
Results:	10247

E.Coli Proteome	
Query:	`length:[50 TO *] AND organism:"Escherichia coli (strain K12) [83333]" AND proteome:up000000625`
Results:	4227

Human Proteome	
Query:	`length:[50 TO *] AND organism:"Homo sapiens (Human) [9606]" AND proteome:up000005640`
Results:	65894

Type IIp Restriction Enzymes with Recognition Sites	
Downloaded REBASE from `http://rebase.neb.com` and selected typeIIp	
Results:	18664

A.2 Motif Extraction and Visualization

To visualize the motifs from the trained DCN models, we adapted the methods of [25] to the case of protein sequences. We first generate a Position Frequency Matrix (PFM) derived from each filter's activations in the convolutional layers. In particular, we score all the input protein sequences using the convolutional, rectification and max-pooling stages of CoMET, and subsequently align the parts of the sequences with a score that passed the activation threshold (> 0) for each filter. After the alignment, we generate the PFM and the sequence logo using the *Weblogo*3.5 python package.

Bibliography

[1] Kristian Stromgaard, Povl Krogsgaard-Larsen, and Ulf Madsen. *Textbook of Drug Design and Discovery, Fourth Edition.* CRC Press, October 2009.

[2] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, March 2012.

[3] K Kahn and K W Plaxco. Principles of biomolecular recognition. *Recognition Receptors in Biosensors*, 2010.

[4] J A McCammon. Theory of biomolecular recognition. *Current Opinion in Structural Biology*, 8(2):245–249, April 1998.

[5] Irwin D Kuntz. Structure-Based Strategies for Drug Design and Discovery. *Science*, 257(5073):1078–1082, August 1992.

[6] Rachelle J Bienstock. Computational Drug Design Targeting Protein-Protein Interactions. *Current Drug Metabolism*, 18(9):1240–1254, 2012.

[7] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, Nezar A Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTOObesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med*, 373(10):895–907, September 2015.

[8] Vikki G Nolan, Clinton Baldwin, Qianli Ma, Diego F Wyszynski, Yvonne Amirault, John J Farrell, Alice Bisbee, Stephen H Embury, Lindsay A Farrer, and Martin H Steinberg. Association of single nucleotide polymorphisms in klotho with priapism in sickle cell anaemia. *British Journal of Haematology*, 128(2):266–272, January 2005.

[9] Thomas J Cradick, Eli J Fine, Christopher J Antico, and Gang Bao. CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Research*, 41(20):9584–9592, November 2013.

[10] Vikram Pattanayak, Cherie L Ramirez, J Keith Joung, and David R Liu. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nature methods*, 8(9):765–770, September 2011.

[11] Shengdar Q Tsai, Nicolas Wyvekens, Cyd Khayter, Jennifer A Foden, Vishal Thapar, Deepak Reyon, Mathew J Goodwin, Martin J Aryee, and J Keith Joung. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature biotechnology*, 32(6):569–576, June 2014.

[12] Benjamin P Kleinstiver, Michelle S Prew, Shengdar Q Tsai, Ved V Topkar, Nhu T Nguyen, Zongli Zheng, Andrew P W Gonzales, Zhuyun Li, Randall T Peterson, Jing-Ruey Joanna Yeh, Martin J Aryee, and J Keith Joung. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, 523(7561):481–485, July 2015.

[13] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–85, January 2016.

[14] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

[15] Daniel Schwartz and Steven P Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology*, 23(11):1391–1398, November 2005.

[16] Asa Ben-Hur and Douglas Brutlag. Sequence Motifs: Highly Predictive Features of Protein Function. In *Feature Extraction*, pages 625–645. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[17] Asa Ben-Hur and Douglas Brutlag. Remote homology detection: a motif based approach. *Bioinformatics (Oxford, England)*, 19(suppl 1):i26–i33, July 2003.

[18] Pål Puntervoll, Rune Linding, Christine Gemünd, Sophie Chabanis-Davidson, Morten Mattingsdal, Scott Cameron, David M A Martin, Gabriele Ausiello, Barbara Brannetti, Anna Costantini, Fabrizio Ferrè, Vincenza Maselli, Allegra Via, Gianni Cesareni, Francesca Diella, Giulio Superti-Furga, Lucjan Wyrwicz, Chenna Ramu, Caroline McGuigan, Rambabu Gudavalli, Ivica Letunic, Peer Bork, Leszek Rychlewski, Bernhard Küster, Manuela Helmer-Citterich, William N Hunter, Rein Aasland, and Toby J Gibson. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, 31(13):3625–3630, July 2003.

[19] Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian J A Sigrist. The PROSITE database. *Nucleic Acids Research*, 34(Database issue):D227–D230, January 2006.

[20] John C Obenauer, Lewis C Cantley, and Michael B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13):3635–3641, July 2003.

[21] Kazuhito Shida. GibbsST: a Gibbs sampling method for motif discovery with enhanced resistance to local optima. *BMC bioinformatics*, 7(1):486–18, 2006.

[22] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

[23] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649, 2013.

[24] A Krizhevsky, I Sutskever, and G E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural ...*, 2012.

[25] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, August 2015.

[26] Aravindh Mahendran and Andrea Vedaldi. Understanding Deep Image Representations by Inverting Them. November 2014.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv.org*, December 2013.

[28] A Dosovitskiy and J Tobias Springenberg. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE ...*, pages 1538–1546. IEEE, 2015.

[29] Jifeng Dai, Yang Lu, and Ying-Nian Wu. Generative Modeling of Convolutional Neural Networks. December 2014.

[30] Ehsaneddin Asgari and Mohammad R K Mofrad. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS ONE*, 10(11):e0141287, November 2015.

[31] Iain Melvin, Jason Weston, William Stafford Noble, and Christina Leslie. Detecting Remote Evolutionary Relationships among Proteins by Large-Scale Semantic Embedding. *PLoS Computational Biology*, 7(1):e1001047, January 2011.

[32] V Nair and G E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International ...*, 2010.

[33] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 52–59. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[34] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. December 2014.

[35] Timothy Dozat. Incorporating nesterov momentum into adam. *Technical Report*, 2016.

[36] Lutz Prechelt. Early Stopping — But When? In *Neural Networks: Tricks of the Trade*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[37] James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[38] L Maaten and G Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[39] Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao, and Feng Zhang. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, 31(9):827–832, September 2013.

[40] W Arber. Host-controlled modification of bacteriophage. *Annual Reviews in Microbiology*, 1965.

[41] Addison V Wright, Samuel H Sternberg, David W Taylor, Brett T Staahl, Jorge A Bardales, Jack E Kornfeld, and Jennifer A Doudna. Rational design of a split-Cas9 enzyme complex. *Proceedings of the National Academy of Sciences of the United States of America*, 112(10):2984–2989, March 2015.

[42] Morgan L Maeder and Charles A Gersbach. Genome-editing Technologies for Gene and Cell Therapy. *Molecular Therapy*, 24(3):430–446, March 2016.

[43] Jeffrey C Miller, Siyuan Tan, Guijuan Qiao, Kyle A Barlow, Jianbin Wang, Danny F Xia, Xiangdong Meng, David E Paschon, Elo Leung, Sarah J Hinkley, Gladys P Dulay, Kevin L Hua, Irina Ankoudinova, Gregory J Cost, Fyodor D Urnov, H Steve Zhang, Michael C Holmes, Lei Zhang, Philip D Gregory, and Edward J Rebar. A TALE nuclease architecture for efficient genome editing. *Nature Biotechnology*, 29(2):143–148, February 2011.

[44] Summer B Thyme, Sandrine J S Boissel, S Arshiya Quadri, Tony Nolan, Dean A Baker, Rachel U Park, Lara Kusak, Justin Ashworth, and David Baker. Reprogramming homing endonuclease specificity through computational design and directed evolution. *Nucleic Acids Research*, 42(4):2564–2576, February 2014.

[45] Alfred Pingoud, Geoffrey G Wilson, and Wolfgang Wende. Type II restriction endonucleases–a historical perspective and more. *Nucleic Acids Research*, 42(12):7489–7527, July 2014.

[46] R D Morgan and Y A Luyten. Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. *Nucleic Acids Research*, 37(15):5222–5233, August 2009.

[47] Richard J Roberts, Marlene Belfort, Timothy Bestor, Ashok S Bhagwat, Thomas A Bickle, Jurate Bitinaite, Robert M Blumenthal, Sergey Kh Degtyarev, David T F Dryden, Kevin Dybvig, Keith Firman, Elizaveta S Gromova, Richard I Gumport, Stephen E Halford, Stanley Hattman, Joseph Heitman, David P Hornby, Arvydas Janulaitis, Albert Jeltsch, Jytte Josephsen, Antal Kiss, Todd R Klaenhammer, Ichizo Kobayashi, Huimin Kong, Detlev H Krüger, Sanford Lacks, Martin G Marinus, Michiko Miyahara, Richard D Morgan, Noreen E Murray, Valakunja Nagaraja, Andrzej Piekarowicz, Alfred Pingoud, Elisabeth Raleigh, Desirazu N Rao, Norbert Reich, Vladimir E Repin, Eric U Selker, Pang-Chui Shaw, Daniel C Stein, Barry L Stoddard, Waclaw Szybalski, Thomas A Trautner, James L Van Etten, Jorge M B Vitor, Geoffrey G Wilson, and Shuang-yong Xu. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Research*, 31(7):1805–1812, April 2003.