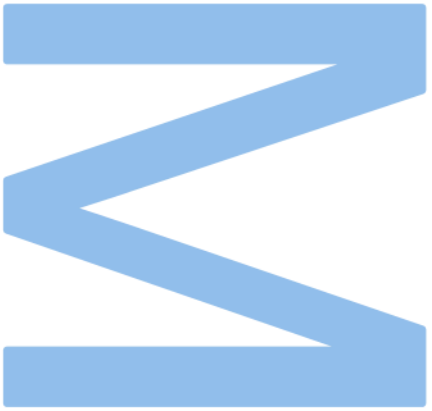

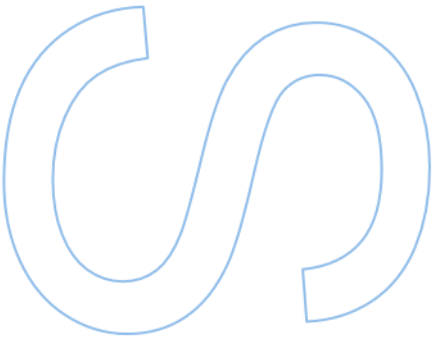

Rafael Vieira,
Master's in Bioinformatic and Computational Biology,
Department of Computer Sciences,
Faculty of Sciences of the University of Porto
2023/2024

Designing In-Silico Aptamers for Potential Use in Marine Bioremediation

Rafael Vieira,
Dissertation carried out as part of the Master's in Bioinformatic and Computational Biology
Department of Computer Sciences,
2023/2024

Supervisor
Sérgio F. Sousa, BioSIM

Co-supervisor
João Carneiro,CIIMAR

Co-supervisor
Diogo Pratas, Universidade de Aveiro

Acknowledgements

I would like to begin by thanking my supervisors, Professor Sergio Sousa, Professor João Carneiro and Professor Diogo Pratas, for guiding me and helping me grow during the development of this work. I am also thankful to my colleagues for sharing this journey with me.

I would like to thank my closest friends - Diogo, Joao, Mariana and Nelson - alongside my little sister, for ever so patiently listening to me talk about a field outside of their own.

Lastly, I would like to thank Wedsy for supporting me throughout all this process. It’s unimaginable to ponder the development of this project without her by my side.

Resumo

Aptâmeros são pequenas sequências de nucleótidos com função catalítica que se apresentam como alternativas promissoras ao uso de antígenos, devido a sua maleabilidade química. Tradicionalmente, os aptâmeros são obtidos através de SELEX, um processo experimental cíclico no qual sequências aleatórias de nucleótidos são expostas a um componente alvo. As sequências que demonstram afinidade para com o alvo, são amplificadas após cada exposição. Apesar de funcional, esta metodologia pode se tornar dispendiosa em termos de tempo, materiais e conhecimento; o que pode impactar negativamente o acesso a este campo de pesquisa. Apesar de ser uma área promissora, ainda é muito recente e carece de métodos ou plataformas estandardizadas de divulgação de informação. Já houve várias tentativas de criação de uma base de dados para divulgação de aptâmeros, mas esses desenvolvimentos paralelos levaram a uma grande dispersão da informação, o que influencia a pesquisa preliminar de interações aptâmero-alvo. Em conjunto com a falta de estandardização de métodos de publicação, a natureza do protocolo SELEX, também impacta negativamente o desenvolvimento de conhecimento de caráter´ısticas por detrás da interação entre aptâmero e alvo. Sendo que publicações frequentemente enfatizam o protocolo e possíveis aplicações do aptâmero, negligenciando detalhes da interação deste com o alvo. Considerando estas dificuldades, neste trabalho desenvolveu-se uma base de dados - AptaCom (https://jc-biotechaiteam.com/AptaCom/) - com o objetivo de conjugar a informac¸ao de aptâmeros dispersa por várias fontes; e foi criado um modelo de 'machine learning'. Este último permitiu aprofundar o conhecimento sobre as interações aptâmero-alvo, sendo também aplicável como uma ferramenta de pesquisa preliminar. Esta premissa foi explorada através da aplicação do modelo no contexto de potencial biorremediação. Desta forma, este trabalho descreve mais um passo para o desenvolvimento da área de pesquisa de aptâmeros, tendo resultado em novos conhecimentos e ferramentas para o efeito.

Palavras-chave: Aptâmeros,Base de dados,Biorremediação,Modelação *in-silico*, Machine learning

Abstract

Aptamers are small single stranded nucleotides with catalytic activity; promising alternatives to usage of antigens, due to their chemical malleability. Traditionally, aptamers are obtained through an experimental process called SELEX, which exposes a randomized pool of nucleotides to a target, and iteratively amplifies sequences that bind with it. Despite its functionality, it is a demanding process requiring not only time, but also costly materials and expertise. These setbacks can make it difficult to reach the entry level of this area of research. Despite the promise of this field, it is still very recent and lacks conventional standardized protocols and platforms for information publication. There have been various attempts of creating an aptamer database, but these parallel developments lead to extreme information dispersal, which can negatively impact preliminary stages of research. Beyond the lack of a standardized approach for information publication, the iterative nature of the SELEX protocol also impacts the understanding of aptamer-target interaction. Published works tend to emphasize the protocol and applicability of aptamers, often neglecting the critical components that describe their interaction with a given target. Considering these issues that plague the field of aptamer research, in this work we developed an aptamer database - AptaCom (https://jc-biotechaiteam.com/AptaCom/) - which bridges information from different sources into one unified platform; and created a machine learning model. The latter allowed deepening the understanding of aptamer target interaction, while also serving as a tool for preliminary aptamer research. This premise was tested through experimentation in the context of potential marine bioremediation. As such, this work serves as another stepping stone for the field of aptamer research, bringing upon new knowledge and new infrastructure.

Keywords:Aptamers,Database,Bioremediation,*in-silico* modeling,Machine learning

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Fcup	Faculty of Sciences of the University of Porto
Up	University of Porto
API	Application Programming Interface
AIR	Ambiguous Interaction Restraints
AN	Nº of total possible amino acid nucleotide interaction
AIF	Absolute Interaction Frequency
ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
CP	Correct Predictions
FP	False Positives
FN	False Negatives
FAIR	Findability Accessibility Interoperability Reusability
Fan	Frequency of amino acid nucleotide interactions
GB	Gradient Boost
IFS	Incremental Feature Selection
mRMR	Minimum Redundancy Maximum Relevance
MAE	Mean Absolute Error
NAKB	Nucleic Acid Knowledge Bank
PDB	Protein Data Bank
RIF	Relative Interaction Frequency
RF	Random Forests
SVM	Support Vector Machine
ToP	Total number of Predictions
TP	True Positives
VT	Variance Threshold
XGB	XGBoost

1. Preface

1.1. Background

Aptamers - introduced in 1990 - can be described as oligonucleotide sequences between 50 and 150 units, that are sequentially selected to bind towards specific targets (i.e., small molecules, heavy metals, proteins, etc). Their specificity, affinity, and chemical malleability make them functional biomaterials in the context of biosensors or target drug delivery systems.

Ever since the inception of the field, there has been a pressing need for an information hub that compiles standardized aptamer information. However, only some could maintain and adapt to the field's growth, resulting in many of them being inactive. With the recent increase of interest in the field of aptamer research, there have been new attempts to develop databases that compile and standardize aptamer information, which results in the overall information being dispersed between different sources, each emphasizing different aspects of the research. This data incoherence and dispersal makes information more difficult to access and compare, which are critical aspects of preliminary research. These setbacks call for a standardized method for bridging this information.

As a result of the process of selection of aptamers - Selective Evolution of Ligands by Exponential Enrichment (SELEX) methodology - there always has been an interest in employing machine learning techniques to improve the experimental pipeline. Such models could further the understanding behind their structure, while bridging the existing accessibility gap for aptamer research. This could allow researchers to confidently explore new possibilities before advancing with an experimental component.

In the past ten years, various attempts have been made to model aptamer target interactions, these attempts were often negatively influenced by the lack of standardized experimental information. Initial models tended to emphasize structural characteristics of both aptamer and target sequences (i.e. pseudo amino acid count, k-Mers etc), often neglecting other features - specially three-dimensional features. In addition, previous models often relied on datasets with poorly described targets - target

description did not include FASTA sequences or even external ID's for cross-referencing. Exploring these often overlooked characteristics, could prove useful for understanding their part in aptamer-target interactions, while also improving the expected performance of a characterizing machine learning model. To further understand model applicability in the context of aptamer research, exploring a potential target to employ the model may return promising results. In the context of this work, focusing on targets such as potential pollutants, showcases the application of this research for future developments in the field of bioremediation.

1.2. Goal

This project has three primary goals: first, to develop a comprehensive database that integrates and standardizes data from a wide range of both active and inactive sources; second, to construct a machine learning model capable of predicting the interactions between an aptamer and a protein target; third, to evaluate the model's applicability, in the context of marine bioremediation.

1.3. Risk Assessment and Impact

When working in the field of bioinformatics, the key risk is its interdisciplinary aspect. This work bridges various areas of research, but mainly experimental and computational biochemistry, machine-learning and environmental sciences. These broad fields are too complex to be mastered and fully explored within the time-frame of this work. To combat this difficulty, this work emphasized mostly practical experimental information, which was more directly translatable into modeling.

Aptamer research shows great promise, with its applications including medical and pharmacological research, environmental sciences and broad spectrum biotechnology research. Despite possible setbacks, this research displays promises of a positive impact, as it brings new tools and publications that may further promote interest in this growing field.

1.4. Hypothesis

The use of machine learning models in aptamer research not only provides practical benefits but also allows for a better understanding of the underlying mechanisms. This work is guided by three main hypotheses. First, using a comprehensive database will lead to the creation of a stronger and more effective model. Second, including features that describe both the surfaces of the protein target and the structural interaction between the protein and the aptamer will improve the model's performance. Third, such a model can be applied in the context of bioremediation by correctly identifying if an aptamer-target pair is valid or not. This validation can then be employed in the process of biosensor development.

1.5. Contributions

The development of this work brought various tools and pipelines for different aspects of its related research. Firstly, the development of the database AptaCom - accessible through https://jc-biotechaiteam.com/AptaCom/ - for a broad reaching aptamer research. With the goal of analyzing and understanding aptamer-target interaction, a tool for mapping their docked complex was developed, along with a metric: Relative Interaction Frequency. Finally, as was the goal of this research, it resulted in a machine learning model capable of identifying aptamer-protein binding, as well as an implementation of this model. These tools can be downloaded in their entirety through https://github.com/rpgv/AptaCom/tree/main. Beyond the tools themselves, this thesis resulted in the publication and presentation of 4 posters in the following events:

• Bioinformatic Open Days (BOD);

• Young Researchers Meeting of the University of Porto;

• Science Meeting 2024;

• BTC;

Posters available here: *https://www.researchgate.net/profile/Rafael-Vieira-27*

1.6. Thesis Structure

The following work is divided into two themes, database creation and model development. The thesis is composed of a preface, introduction, methodology, results, discussion and conclusion. Each section of this work is partially sub-divided into the two overarching themes.

2. Introduction

The following section describes the theoretical background for this work, covering the appearance of aptamer and SELEX technologies, as well as the history of *in-silico* aptamer modeling.

2.1. Aptamers

Aptamers are short single-stranded oligonucleotides made of RNA, DNA, or modified nucleotides, typically consisting of 50 to 150 units. These bio-materials have the ability to fold into a diverse array of shapes, similar to how proteins fold into secondary and tertiary structures. This allows aptamers to bind to various molecular targets, from small molecules to proteins.

The presence of phosphate in the aptamer's backbone renders it negatively charged, influencing its interaction with targets, particularly proteins. The structure of an aptamer is typically based on a pre-existing template with conserved regions; a random pool of nucleotides is arranged in relation to those regions during development. This process dictates how the modified template will bind to a specific target. Aptamers can adopt secondary structures, such as triplexes, quadruplexes, hairpins, bulges, pseudoknots, stems, and loops. These diverse structural motifs contribute to the versatility of aptamers in molecular recognition and binding

2.2. Aptamer Discovery

The term ”aptamer” was coined in 1990 in the publication titled ”In vitro selection of RNA molecules that bind specific ligands” by Andrew D. Ellington and Jack W. Szostak [1]. In their research, they aimed to explore whether early life could have originated from random polymer sequences and to determine the likelihood of these structures developing catalytic activity. To assess this catalytic activity, they synthesized a library of randomized RNA sequences and amplified them through PCR (polymerase chain reaction). The amplified library was then subjected to various consecutive rounds of affinity chromatography, individually targeting a set of six organic dyes. After each round of chromatography, the binding structures were amplified. Their findings indicated that around 50% of the sequences demonstrated specific binding after five to six rounds of chromatography and amplification

2.3. SELEX

The Systematic Evolution of Ligands by Exponential Enrichment (SELEX) was introduced in 1990 by Tuerk and Gold [2]. Their work outlines the method for selecting high-affinity nucleic acid ligands for a bacteriophage polymerase. The approach is based on an evolutionary perspective, involving ”variation, selection, and replication,” which translates to library creation, incubation, and PCR amplification. The library was created based on a template sequence that contained a randomized ’hairpin loop’ section, resulting in an extensive library of oligonucleotides. The incubation of the oligonucleotides and their target lasted 4 minutes at 37°C, followed by filtration using nitrocellulose filters. The bound RNA was recovered from the filters, and complementary DNA copies were made through PCR amplification.

Since its introduction, there have been various iterations and modifications of the SELEX methodology, all following the initial premises of ’variation, selection, and amplification.’ Some examples include capillary electrophoresis SELEX, magnetic beads SELEX, and whole-cell SELEX [3]. These modifications have improved the technique, whether in the bound sequences’ recovery process or the process’s automation.

2.4. Aptamer-Protein interaction

Aptamer protein interaction can be categorized into nucleotide-binding proteins and protein-binding nucleotides. In the first case, the driver of the interaction is the protein - usually an enzyme - that has an active site that fits its nucleotide target; as such, the selectivity comes from the protein and not necessarily the nucleotide. This type of behavior can be observed in instances of gene regulation. On the other hand, in the case of protein binding nucleotides like aptamers, the selectivity of the process is driven by the aptamer itself; it allows the aptamer to bind to any type of protein, not just enzymes. This interaction can be described primarily by their opposing chargers. Aptamers have a negatively charged backbone, predisposing them to interact with positively charged molecules or protein surface areas. It is also common to observe aptamer interactions between proteins that naturally bind to nucleotides [4]. This process usually takes place along the binding sites of the protein, as the aptamer shape aligns with that of the protein target.

Like enzymes, aptamer-protein binding is heavily influenced by it's three-dimensional shape. Besides aptamer backbone charges, their shapes are essential for various interactions. Each conformation can relate to a degree of more or less nucleotide exposure. In the case of exposed nucleotides, bonds, such as hydrogen bonds, can be more frequent - depending on the exposed nucleotide. At the same time, structures with less or no nucleotides exposed will rely more heavily on charged interactions.

2.5. The data issue

Since the discovery of the aptamer, there is yet to be developed a database that coordinates the aptamer information in a similar way as the Protein Data Bank organizes protein information. There have been various attempts to develop a complete database [5], [6]. However, due to oscillation of interest in and relevance of the field throughout the years, most were either deactivated or are no longer updated. With the renewed interest in the field in recent years, new attempts have been made to build a consistent aptamer database [7], [8], [9]. These efforts are happening simultaneously across different groups, both academic and industry-focused. As a result of this parallel database development, each database takes a different approach and emphasizes different features of the SELEX pipeline, the aptamer and its target. In addition, this isolates the information and associated tools of each database, which can complicate more global preliminary research. This growth points to a need for a more robust information infrastructure aptamer research, also mentioned by Miller et al [10]. There are promising databases, both in structure and standardization techniques (i.e. Nucleotide Knowledge Bank), but these often contain only a reduced number of aptamer samples. In-silico modeling of biochemical interactions is a computationally intensive and information demanding process. It requires extensive understanding of the molecules at play, often conveyed through their PDB file, which describes each molecule in terms of their atomic composition and consequential three-dimensional structure. With current developments in the field of machine learning and in-silico modeling combined with the lack of a standardized infrastructure, a data availability issue arises. In aptamer research, the focus is usually on the functionality and applicability of the aptamer towards a specific target, with little emphasis on how the complex is formed and how the two components interact. It is common for aptamer publications to describe both aptamer sequence and target superficially, with often no reference or analysis of their final complex. These restrictions, along with the few

available active aptamer databases, culminate in the aforementioned data availability issue. Of the few accessible databases, most lack a description of the aptamer-target complex that can be translated into in-silico modeling, with features such as target sequence often missing. Another issue that plagues this field regards data reproducibility and validation. A review [10] analyzed 780 articles, and observed that approximately 41% omitted explanations on sequence mutations, which can greatly difficult reproducibility later on. Combined with the fact that available databases tend to focus more on data compilation than verification, it raises the question of reliability of available information. This is especially aggravated by the data dispersal caused by the parallel development of multiple unrelated databases, which can be challenging to cross-reference.

Despite isolated progress in the development of a standardized approach to retrieval, verification and publishing of aptamer data, information remains dispersed and diverse. There is a need to develop an information hub analogous to PDB. By creating a similar infrastructure, it becomes easier to enforce a more standardized approach of the data to be published, and to perform a more direct comparison when researching possible aptamer-target pairs

2.6. Databases

The following section provides an overview of the various databases utilized in this master thesis study. Each database offers unique strengths and limitations, reflecting the diversity of approaches in aptamer research. Their relevance led to the selection of these databases based on aptamer-target interactions, data accessibility, and the extent of detailed information they provide.

2.6.1. Nucleotide Knowledge Bank

Borrowing from PDB's data structure, the Nucleotide Knowledge Bank (NAKB) [11] compiles extensive information from a large array of nucleotides. It represents one of the best examples available to the research community in terms of structure and standardization. Each entry describes the nucleotide (singular or in a complex with a target) in terms of their publication title and reference, as well as related key-words that describe it. Finally, each entry relates back to their PDB ID, where further information regarding sequence and three-dimensional structures can be found. However, from the 18 000 plus entries it possesses, only a few describe aptamers (approximately 150),

fewer still represent protein binding aptamers, which are often the focus of modeling studies. Additionally, its feature set does not contain experimental information such as binding affinity of the complex.

Despite its promising data structure and information richness, it is still not applicable as a standard in the context of aptamer research.

2.6.2. Aptamer base by Freebase

One of the most cited databases in the field, and hosted by the now inactive Freebase [6] relational database, Aptamer base was crucial for preliminary aptamer modeling. It contains approximately 4500 entries, describing a diverse set of aptamer-target interactions. These 4500 entries are, to some extent, redundant as they describe different assays of the same 725 aptamer-target pairs. In 2016 the Freebase platform, which hosted Aptamer base, was bought by Google and later deactivated; however, there are still references on github to its aptamer focused dataset [12] that allow relatively easy access. This dataset suffers from the aforementioned target description issue, in which targets are only broadly referenced. In the case of protein targets, there are no references other than a name and what it is thought to be internal ID. The internal ID could in theory link to a more detailed description, however that data is no longer available for confirmation. Unlike most databases, Aptamerbase has a 'validation' descriptor, which relies on its 'redundant' assay based structure, to describe entries as 'validated by' other publications.

2.6.3. AptaIndex by Aptagen

Aptagen develops and commercializes aptamers for research at the academic and industry levels. Their database - AptaIndex [13] - serves as an all-in-one system that allows researchers to search and order aptamers in one single platform. With approximately 800 entries, AptaIndex describes aptamer in terms of its chemistry and sequence, some degree of experimental conditions (i.e. binding buffer, and protocol applied) and the target solely by name, following the trend. Although this database does not guarantee validation of all their entries, they offer a validation service, for an additional fee. It does not solve the broader problem, but it offers a way to combat it in a preliminary research phase

2.6.4. Aptabase by IITG

With the increasing relevance of the field of aptamer research, some academic institutions created projects that promoted the development of new databases, through an extensive manual reviewing process. Aptabase [9] is one such example. Developed by the Indian Institute of Technology of Guwahati, Aptabase is described as a potential hub for collecting aptamer information world wide. The database is hosted by the institution itself and contains approximately 600 entries. It describes the aptamer in terms of published name and sequence, aptamer-target interaction in terms of its affinity, and target solely by name. The database does not guarantees validation of the entries

2.6.5. UTexas Aptamer Database by University of Texas

UTexas aptamer database [8] represents this group's second attempt in developing an aptamer database. There is not much information available on its predecessor [5] - published in 2004 - but the current database - published in 2023 - aims to standardize aptamer information while combating the aforementioned reproducibility issue [10]. The research group responsible developed a whole academic course, focusing on research and review process, to analyze publications on aptamers that had a complete description of the aptamer sequence modifications. The resulting database covers 1400 entries, covering features analogous to the previous databases, following the same pattern of target description. Despite the focus on sequence verification, the published data is not externally validated

2.6.6. AptaDB by LMMD

Developed by the Laboratory of Molecular Modeling and Design School of Pharmacy, and the East China University of Science and Technology, AptaDB [7] comprises 1300 entries of aptamer-target interaction. Unlike any other currently available academic database, AptaDB builds upon the previously described features, and links each target to an external database with more detailed information. Every protein target is linked to Uniprot, while small-molecules are linked to PubChem. The overall structure of the database is very cohesive and complete, and it sets a good precedent for a potentially standardized structure. Additionally, the detailed description of targets shows promise for applications in in-silico modeling.

2.7. Aptamer *in-silico* modeling

Aptamer in-silico modeling is a new field of research that promises an alternative to the traditional SELEX methodology. Experimental processes as a whole tend to be costly in terms of time, labor and material expenses. The employment of such computational approaches can accelerate research while reducing costs. The following section describes an overview of the history of aptamer in-silico development, the key models, their abilities and limitations

2.8. History

Aptamers are versatile biomaterials, with a scope similar to that of antibodies, while being easier to modify with less compromise in terms of chemical function. However, the SELEX methodology brings some setbacks to aptamer development, namely cost in materials, maintainability and time. Considering those factors, the interest in developing a machine learning model in aptamer research is clear. Since the discovery of aptamers, there have been various attempts to develop a machine learning model that could accelerate the aptamer selection process. Starting around 2012, there was a visible increase in published research and models in this field [14]. Following this increase in interest, in 2014 Li. et al [15] developed one of the first well known machine learning models for aptamer-target characterization, with emphasis in protein targets. In their work, they employed a Random Forest predictor, performed feature selection and posterior analysis to understand aptamer-target interactions further. Their final feature set consisted of 220 features - selected through incremental feature selection (IFS) and minimum redundancy maximum relevance (mRMR). It emphasized structural features, namely k-mer and pseudo-amino acid composition. Their resulting model showed promising results, and set a precedent to the field. The work performed by Li et al. served as a key basis for various classifiers later developed, such as Yang et al. [16], who explored a similar dataset, but with a new feature space. More recently, works like AptaNet [17] employed deep learning techniques with the goal of classifying aptamer-target interactions. This novel approach to in-silico aptamer development showed promising results within their benchmark dataset

2.9. The dataset issue

As previously mentioned, there are various available databases. However, only some contain detailed information applicable to machine learning due to the often poor target description - specifically protein target description. Only recently, with the appearance of the AptaDB database, there was a detailed connection between aptamer and protein target, allowing for an analysis that is, in principle, more coherent with the biochemical interactions displayed in a non-simulated environment. To circumvent the absence of such a database, Li et al. used Aptamer base's [6] target information, which only described the target's name. They performed a name match search through the swiss-prot database, saving only entries representing a complete name match. This protocol returned approximately 725 entries with approximately 170 protein targets, which were later expanded with negative entries. The negative dataset was created by randomly matching aptamer and protein targets from the positive sample while guaranteeing no overlap. Even though the work of Li et al. returned promising results, the existence of possible incongruities in their dataset could have rendered the model weaker with external samples. These setbacks are a testimony of the lack of available data 10 years ago (2014).

2.10. Recent developments

More recently in 2023, google launched its own aptamer model - AptaBERT [18] - which takes a different approach to aptamer-target interaction analysis, by employing a language model - BERT. Bidirectional Encoder Representation from Transformers (or BERT), was published in 2018 as a versatile transformer, capable of being easily adapted to an array of NLP (Natural Language Processing) related tasks. As such, with a large enough dataset, it can be an effective model for approaching sequence data, such as aptamer and target sequence.

One key setback to aptamer modeling is the need for confirmed interaction data. Of the little data available, it is often hard to reproduce, and it would be both expensive and time-consuming for researchers to attempt to confirm each entry in terms of precision. To circumvent these issues, the researchers resorted to a large private dataset owned by a pharmaceutical company - Dianox - that compiles more than 67.000 validated aptamer-target interactions. A dataset of that magnitude is vital for models such as

BERT. Additionally, it endows their research with more flexibility and precision for a deeper understanding of aptamer target interactions.

2.11. Tools

Advancements of computational techniques have greatly improved biological simulations. These tools employ internal models and algorithms to either extract features of biological components, or to simulate the interaction of a pair of molecules. In the following section, some of the key tools employed in this work are briefly described in terms of their function, and user interface.

2.11.1 Unafold

Released in 2008, UNAFold [19] is a software tool set that allows simulation of various aspects of single-stranded nucleic acid sequences, namely: hybridization, melting pathways, or secondary structure. Its suite of tools can be downloaded and run on linux/unix systems locally via command line interface or used through their web server. It takes as input the sequence and its predictions can be obtained through various parameterized options (i.e. with or without computation of base pair probabilities). Which in turn allows to prioritize either precision or computational demand during folding

2.11.2 3DRNA/DNA

3DRNA/DNA [20] [21] is a software tool that simulates the three-dimensional folding of linear (or circular) strands of RNA and DNA. It takes a few use case specific parameters, the nucleotide sequence and its secondary structure. Then it deconstructs the secondary structure into a graph that organizes the secondary structures consecutively. It then iterates through each component of the secondary structure and searches for possible matching three-dimensional templates - experimentally obtained. In case of absence of a match, the program employs a Distance-Geometry or bi-residue method in an attempt to generate three-dimensional templates. Assembly is done by joining each matching 3D component to its parent node. After assembly, it is possible to add an optimization step in which 3DRNA uses Simulated Annealing Monte Carlo optimization (SAMC). This tool can be accessed via their web portal or by downloading and installing.

2.11.3 Docking

Molecular docking (or Docking) is a computational methodology that allows for simulation of interaction in between two molecular components at the atomic level. Most commonly, docking is employed in the context of drug research and development. It takes detailed three dimensional structures - usually PDB file format - as input, and simulates various possible structures for the ligand within the receptor's selected site. After simulation, the refinement phase begins, and structures are narrowed down, based on the overall complex stability. Molecular docking has played a key part in bridging access to small scale and preliminary assays, without the requirement of laboratory materials; while also allowing for large scale exploration of potential targets, with lower cost. Docking development usually focuses on protein-ligand interactions; however in the past two decades there has been a surge in interest in RNA/DNA-protein docking [22]. Nucleotide-protein complexes cover both protein-RNA/DNA and DNA/RNA-protein interactions. The first depends on the proteins ability to bind to a certain nucleotide target, and how that interaction is done - i.e. how nucleases bind to their 'target'. The latter describes the opposite, in which the nucleotide does the binding - i.e. ribozymes and aptamers. The increase of interest in this field, brought upon tools such as HADDOCK [23], capable of docking various types of molecules, including nucleotides-protein complexes. HADDOCK differs from most of the available tools, as it allows to flexibly simulate nucleotide-protein interaction as both a receptor and a ligand.

2.11.4 HADDOCK

HADDOCK is one of the most reputable docking tools available, especially regarding nucleotide-protein interactions. Its information driven software suite allows users to complement their computational docking job with experimental data by describing known active sites. It is also capable of performing ab-initio docking by defining the center of mass restraints, and generating a larger sample size of structures. It's process can be broken down into three main steps: rigid-body docking, semi-flexible refinement, and final refinement. Moreover, the whole process can be optimized for bioinformatic prediction data by selecting specific parameters, again relating to the sample size refinement. By default, the final output is a ranked report containing the top 10 clusters with a minimum number of structures defined by the user - default 4. Both

structures and clusters are organized in terms of their ZDOCK score, from best to worst.

2.11.4 SurfMap

SurfMap [24] is an open-source command-line interface tool that enables users to convert PDB protein files into 2D images describing their surface features. The general approach of the tool relies on defining a 'shell' around the protein surface. This 'shell' is composed of units representing the protein residues that are characterized in terms of specific features (i.e. hydrophobicity or electrostatic potential). After characterization, the shell is flattened into a fixed sized grid. The final output consists of a PNG file, containing the 2D image of the protein surface map, and a TXT file, containing the matrix that defines the map.

2.12. Features

A factor commonly neglected in aptamer in-silico modeling is the three-dimensional interaction between the aptamer and the target. Due to factors such as lack of data and the difficulty associated with modeling and extracting three-dimensional information, the critical descriptors employed in these interactions are usually based on sequence information. The two main types of sequence information that can be extracted from biological sequences are structural and physio-chemical. Structural information regards sequence composition - i.e., number of times a given letter or sequence of letters repeats. These techniques, borrowed from natural language processing, allow for an overview of the inner patterns of the sequences to be analyzed. More traditionally, tokens or k-mers are used to describe the relative or absolute frequency of a sequence of k letters of a protein or aptamer sequence. Due to the exponential increase in complexity with the increase of letters, usually around three (triad tokens or 3-mers) are used. Physio-chemical descriptors cover an array of features that correlate the sequence composition to chemical properties like hydrophobicity and charge. More complex features such as autocorrelations [25], i.e. Geary autocorrelation, use specific algorithms to map the distribution of sequence components properties to describe a more detailed profile of sequence composition

2.13. Machine Learning

Machine learning can be described as a set of techniques that use existing data to train a statistical model to predict outcomes not described in the original data source. This

enables a computer to take on specific tasks without a rigid set of commands, relying instead on its task model. Examples of applications range from a simple linear model to represent the price of a house in relation to its area, to the prediction of cancer prognosis [26]. One of the most popular and influential examples of machine learning applied in the field of natural sciences is AlphaFold [27]. This tool is used to predict three-dimensional structures of proteins, which has become vital in the current biomedical research landscape.

The field of the natural sciences is vast and complex, and it is often difficult to understand the intricacies of certain mechanisms. Machine learning techniques can enable a deeper understanding of these complex systems despite not having all the information that describes them.

There are two critical components to machine learning development: the dataset and the model. In classical (supervised) machine learning, the dataset consists of all the information that will be fed to the model, allowing it to approximate the reality of the task at hand to perform predictions. The dataset usually comprises a set of entries, their respective features and the classes they belong to. During development, feature selection or compression processes can occur in which irrelevant features can be removed or similar features can be merged into one. The final dataset should be diverse, comprehensive, and appropriate to the model. Each model has varying advantages and different performances with different datasets and problems. A linear model is constricted to working with linearly correlated data. Alternatively, a more complex model, like artificial neural networks, can better adapt to more varying shapes of data, being, however, more demanding in terms of the amount of data and computational capacity. The learning process (as well as the dataset) is usually divided into two sections: training and testing.

Broadly speaking, the training phase consists of 4 key stages. First, the model is exposed to a set of entries of the training dataset. Based on the features of these entries, the model will attempt to predict an outcome. The prediction will be directly compared with the real value of the entry, and based on the proximity of the prediction with the real value, a certain degree of corrections will be applied. This process is repeated by a certain number of iterations, either previously defined or defined by the model in question. This iterative adaptation of the model allows for identification and

storage of patterns within the dataset. These patterns can then be applied to predict previously unseen cases. To train the model, the available data is usually split into training data and testing data, following an 80% to 20% ratio.

The testing phase can be summed up by the model's exposure to new examples that did not belong to its training dataset. The predictions are again compared with the real values and these results are then used to obtain general metrics of model performance. Traditionally, the metrics used to describe a model are accuracy, precision, recall, F1 and mean absolute error (MAE). Considering correct predictions (CP), total predictions (ToP), true positives (TP), false positives(FP) and false negatives (FN). Accuracy can be defined by the model's ability to predict a correct outcome *(CP/ToP)*100*, while precision describes the amount of correct target predictions *TP/(TP + FP)*. Recall approaches it differently, and measures the number of instances of target classes that are correctly identified *TP/(TP + FN)*. The F1-score uses these past two metrics and calculates a ratio that indicates the model balance of precision and recall *2/((1/Precision)+(1/Recall))*.

K-fold cross-validation is a technique that modifies the initial approach for splitting the dataset; instead of a singular train-test split, it segments the dataset into k parts, subdividing each part into train-test segments following the aforementioned 80% to 20% ratio. This is done to analyze the model behavior while removing the influence of the dataset split. The resulting evaluation metric may vary, but employing the mean value of Mean Absolute Error (MAE) for each fold is common.

In the context of aptamer-target interaction modeling, there are various models applied in the literature, namely: random forests (RF) [28], artificial neural networks (ANN) [29] and support vector machines (SVM) [30]. Famously, RF is one of the first models implemented in the context of aptamer prediction. As the name implies, this tree based algorithm creates randomized decision trees (ensembles), based on random features or bagging - training various models on random sets of data, then joining their predictive ability through averaging or voting. Through this approach, the model selects subsets of the dataset at random for each decision tree of the forest, lowering interrelation between them. This algorithm is flexible and tends to perform well with datasets containing a large number of features, it also tends to perform well when presented with noise, outliers or even missing data. In addition, it attributes an

importance coefficient to the selected features, which can play an important role in the understanding of the model's structure. RF can however, be very memory heavy, and often result in slower prediction times - which could impact future model implementations.

Artificial neural networks (ANN) are inspired by the structure of neurons in the brain, and how those are activated and inhibited when presented with external stimuli. ANN's have a layered structure, segmented in three main sections: input layer, hidden layers and output layer. Each layer is composed of an array of nodes that work as independent functions with their own weights and biases. These functions take an input - which is either the original, or stems from the previous layer - and define an activation threshold that when reached, allow information to pass from the current node to the next. This algorithm is one of the most versatile algorithms available, and is capable of performing complex tasks such as image recognition or DNA sequence compression [31]. Despite its versatility, it requires large datasets and can be very computationally intensive, it also has lower interpretability due to its black box nature.

SVM defines a n-dimensional hyperplane - n is defined by the number of features - that separates the classes to be classified. The classification relies on defining the maximum margin for distinguishing between the most similar points across classes. This model tends to perform well in datasets with smaller sample sizes in relation to the number of features, and tends to be memory efficient. However, SVM's tend to under perform when trained on datasets that have a lot of noise.

XGBoost [32] (XGB) is at its core, a more efficient implementation of Gradient Boosting [33] (GB) , another ensemble model. XGB trains weak models sequentially and with each new model, corrects upon the flaws of the previous. XGB performs well with large datasets, and provides an overview of feature importance through the attribution of feature coefficients. It also strikes an acceptable balance between model simplicity and model performance, which actively helps to reduce overfitting. XGB has also cuda support, allowing it to run on graphics card, greatly improving performance during training.

2.13.1. Model application

Bioremediation can be defined as the application of biological agents in the identification and removal of pollutants.

Water quality assessment is a key component of good public health practices. Waterborne diseases cause approximately 1.6 million yearly casualties worldwide [34]. One of the key causes of contamination in water sources traces back fecal matter, of which one of the most well studied bacteria is the Eschirichia coli (or E.coli). E.coli is a gram-negative bacteria, found in the gastro-intestinal tract of warmed-blooded animals [35]. Due to its public health impact, as well as its deep literature description, E.coli can serve as a promising model for exploring the applicability of the machine learning model developed in this work. E.coli's structure is very well described in the literature, and has prominent protein components such as ompA or ompF - outer membrane protein A and F. These can be targeted as example protein targets to evaluate model applicability.

2.14 Main objectives

The field of aptamer research has been gaining traction in the last decade and has shown great promise in pharmacological research and biosensing. However, this remains a relatively niche field, with gradual but slow adoption by a broader range of research groups. This can be due to the decentralized nature of aptamer data. Alternatively, the additional cost of new infrastructure can be difficult for institutions to enter. These issues segment this work into two primary goals.

2.14.1 Database

To tackle the data diversity and decentralization issue, the aim is to develop a database that serves as a hub of aptamer information that bridges the data from different sources. As such, the goal is to centralize the information in one platform that compiles critical aspects of published aptamer work, while following the FAIR principles of Findability, Accessibility Interoperability and Reusability. This will allow for easier comparison across databases, especially regarding the features made available by each source. This database will not replace the role of existing ones, or the work behind their maintenance. It only aims to ease research across different platforms, guiding researchers to the original database for additional details.

2.14.2 Model

With the goal of tackling accessibility, and to lower the entry level to perform preliminary aptamer research, the aim is to develop a machine learning model, capable of characterizing aptamer-target pairs. The end goal is to not only make aptamer research

more accessible, but to also power a better understanding of aptamer-target interactions. The development of this model will allow the evaluation of performance in relation to added three-dimensional features. This analysis will be done through comparison of multiple models, as to observe their fit with the respective set of data. Hopefully, this will deepen the understanding of aptamer-target interactions, while validating the role of these features in future developments. The last step of development of this model consists in evaluating its applicability in the context of marine bioremediation. This will be done through testing of its ability to correctly identify an aptamer-target pair.

Methodology

This section can be divided into two parts. The first, describes the systematic approaches employed in the retrieval, processing, and utilization of aptamer data across various databases. Given the diverse nature of these databases, both in terms of data availability and the structure of stored information, it was important to adopt a flexible approach to ensure comprehensive data retrieval. The second, emphasizes the methodologies and techniques applied in the development and understanding of a machine learning model for aptamer-target classification. Both aptamers and proteins are very diverse groups of biomolecules, each with intricate components that influence molecular interaction. To properly model these interactions, it was critical to retrieve a comprehensive set of features regarding their composition. To further deepen the understanding of aptamer target interactions, descriptors covering surface protein features, or point of contact between aptamer and target were also developed and evaluated.

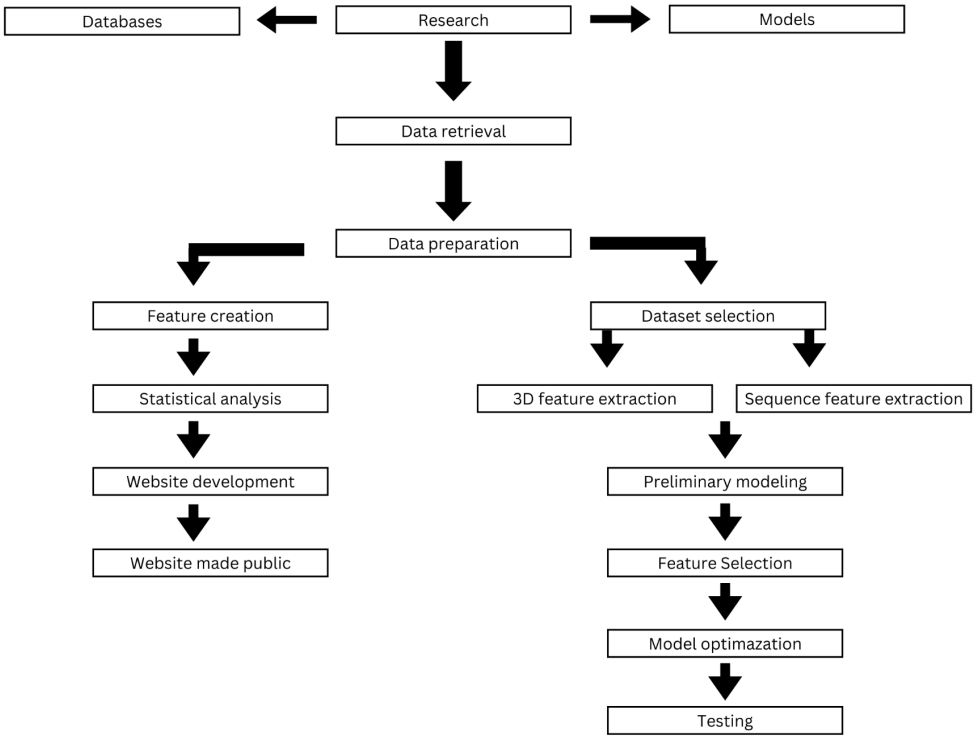

Figure 1 - Workflow followed in the methodology of this work

3.1. Data retrieval

One of the critical issues of the various active aptamer databases is their accessibility, especially in terms of a complete raw dataset. Except for the UTexas Aptamer database and AptamerBase (by Freebase), no other database had all their information available in a file to be used. This means the remaining sources had to be accessed either through web scraping, API keys or a combination of both. The Aptagen database comprises more than 800 entries and was fully accessed through web scraping using the Selenium module for the Python language. This module simulates user access to a given website and allows for information extraction from more dynamic websites. The AptaDB database had a specific section with files available for download; however, those files contained only some of the information available in their dataset. All complementary information, such as binding affinity, was retrieved using selenium. The IITG database was hosted on a simpler website, which allowed a more straightforward approach using Panda's module for Python, which can read tables from an HTML page. The data sourced from the NAKB database was accessed through their API.

It is important to note that each database required an individual approach and related code due to the specificity of web scraping. Approximately 4000 entries were retrieved to various degrees of detail in relation to the information in the respective sources. Considering the two goals of this project, the data was first divided into applicable and nonapplicable for in-silico development. This division was done based on two factors: aptamer targets a protein; the protein target can be traced to a PDB file. Considering that only two sources referenced the aptamer and their corresponding target in detail, data specifically from AptaDB and NAKB was chosen for in-silico development, totaling approximately 630 entries.

3.2. Database creation

Given the different approaches to aptamer research taken by the various databases, there are various degrees of detail and intent behind each of them. The main goal of this database is to reconcile the discrepancies of each database, improving data findability, accessibility, interoperability and reusability (FAIR). This will be done while also promoting a standardized structure. As such, defining relevant features to display

on the database was an important step. The features to display on the database were chosen mainly on their relevance on primary identification of aptamer-target pairs, and in their ubiquity across different databases. As such, in a preliminary research phase, researchers can easily find and compare overlapping features. Subsequently, the chosen features were standardized across the databases regarding data type, sequence representation. Once these features were standardized, the database was integrated into a WordPress website with the WP Table plug-in.

3.3. Docking

To translate some information about how aptamers and protein targets interact in three dimensional space, as well as to evaluate how that information influences a machine learning model, docking across all the positive pairs in the in-silico dataset was performed. To perform docking, it was necessary to first generate the three dimensional structures of the aptamers. To achieve this, the local version of 3DRNA/DNA tool was employed. The samples were first divided into RNA and DNA sequences to create each aptamer's three dimensional structure. Then, a simple bash script - available in the Github repository - was created to loop through the sequences, using the sequences and their secondary structure as input. There were no parameters to further customize in the context of these predictions. HADDOCK is the preferred tool for performing DNA-Protein docking [23]. Preliminary testing showed that each docking took approximately 10 hours, using the default settings for bioinformatics predictions. To combat this issue, docking was performed using a coarse-grained approach [36], also available on the web platform. Coarse-grained reduces the resolution of the docking process by working with a simplified version of the three-dimensional structure of the PDB file. This approach reduced the expected docking time by five times, down to approximately 2 hours per run, without considering queuing or other jobs on the server. Selenium was employed to automate the process, fill the submission forms, and set the parameters for each aptamer-target pair. The web server version of the tool was used due to issues with installation of the local version's dependencies. The parameters employed were coarse-grained molecules - to improve execution time - generating 10.000 structures for the rigid-body stage, 400 for the semi-flexible refinement, and 400 for the last refinement step, as suggested by HADDOCK's own platform in the context of bioinformatic prediction.

Considering the *ab-initio* approach of these samples, docking was performed with ambiguous interactions restraints (AIR's)> 20% - defines random sections of both aptamer and target to dock for each run iteration. The resulting data is automatically ranked in relation to their ZDOCK score and was downloaded via a selenium pipeline.

3.4. Dataset build

After selecting and extracting entries applicable for in-silico modeling based on the presence of a detailed target descriptor, the feature selection process was initiated. The goal of the preliminary dataset was to compile a large number of features and subsequently narrow them down based on feature selection and posterior feature analysis. Protein and nucleotide sequences can be extensively described by their structural and biochemical sequence content. The Python module PyBioMed [37] was used to achieve this. PyBioMed allowed the extraction of features such as unit content (nucleotide or amino acid), tokens (3-kmer and triad amino acid), pseudo amino acid composition, and spatial autocorrelation features. Because the chosen library does not support RNA sequences, all sequential information was converted to DNA instead. This has often been done in the context of aptamer research due to the absence of broader ranging tools [38].

To analyze how the aptamer and the target interact spatially, docking was performed between the aptamer and its target using the HADDOCK web server. The resulting complexes were downloaded and mapped regarding contact points between the aptamer and its target. These points of contact (distance between nucleotide and the amino acid inferior to 5 Angstrom) were stored as two metrics: absolute and relative interaction frequency (AIF and RIF, respectively). RIF can be calculated by the frequency of interaction (Fan) of a Nucleotide (N) with an amino acid (A), divided by the total number of possible contacts between them - Fan/(AN). This feature allowed for some correlation between sequence content and three-dimensional binding information, resulting in a total of 104 columns - 80 (AIF) and 24 (RIF).

The SurfMap [24] tool was employed on the protein targets to bring three-dimensional surface information to the dataset, emphasizing the 'circular-variance' feature - translating a 'topographical' map of the protein surface. This was achieved through standard usage of the tool through the command line, followed by converting its output matrix into a uni-dimensional row. This singular feature was emphasized because each

feature is translated in more than 2500 columns, and other features described in the tool are already described linearly by other components such as auto-correlation and hydrophobicity. The AC2 [39] compression tool was used on the protein target sequences to complement the feature set further. This command-line-based tool returned various metrics, of which dissimilarity rate and Shannon entropy were chosen to be added to the dataset. These features culminate in approximately 12000 columns; this aligned with the small sample set (approximately 630 entries) calls for additional processes, starting with data augmentation. Using negative samples, similar to what was done by Li. et al., allows for expanding the sample size while maintaining a balanced dataset.

The difference lay in analyzing aptamer sequence similarity and combining the targets of aptamers that presented more than 60% sequence similarity into individual groups. Then, the positive dataset was shuffled, and aptamers were paired with targets belonging to aptamers that were at the most 40% similar. This was done assuming that aptamer specificity will guarantee negative examples under these circumstances. A secondary sub dataset was created to compress some possibly redundant information on the dataset. In this dataset, the 20 amino acids were grouped into six categories: positively charged (K, R), negatively charged (D, E), polar uncharged (S, T, C, Y, N, Q, H), aliphatic (G, A, V, L, I, M), aromatic (F, W, Y) and other (P). This grouping only affected the RIF metrics and triad tokens based on groups instead of single amino acids. Despite applying this technique on only two sets of features (RIF and triad tokens), it allowed the conversion of approximately 8000 columns describing protein triad tokens into 216 columns. These features were combined with the existent dataset, into one dataset with some degree of redundancy. This redundant dataset was designed with the goal of observing which features would be chosen by the most promising feature selection approaches.

Regarding missing values, and other outliers, as all structure related features were extracted from one cohesive library, there were no oscillations that required correction. SurfMap features however, were not designed with machine learning in mind, and as such required modifications of ’infinite’ values, which were used to represent ’empty’ grid positions. These infinite values were replaced by 1, as all the remaining values are represented in between 0 and 1 without ever reaching 1, which simplified scaling.

3.5. Model

In the context of this work, the model targets proteins specifically. Due to aptamers wide range of targets, at this point, with current data available, it is preferable to approach targets individually. Each interaction differs extensively, and a model capable of interpreting all possible interactions would require protein targets, make up a large portion of the researched interactions, and cover a broad range of possibly useful targets.

The modeling process, followed mostly an experimental approach, divided into three parts: preliminary modeling; model selection and model optimization.

Considering both the literature and the relatively broad spectrum dataset, the preliminary modeling step focused on exploring different model behaviors, when faced with the created dataset, at this stage, without feature selection. This preliminary stage was developed through the design of an experimental pipeline. Considering information from the literature, there are models highlighted in terms of performance with a feature set similar to the one developed in this work; namely, RF and other ensemble models [15], [16], as well as artificial neural networks (ANN) [17]. The experimental pipeline can be summed up as an array of models that were iteratively trained on the dataset: RF, ANN, XGBoost, and SVM. Scaling was taken into account at this stage, with standard sci-kit learn scaler being used.

After selecting the most promising model, the preliminary exploration continued with testing a model structure: conventional singular model versus stacked ensemble model. The model being developed aims to prevent researchers from requiring computationally intensive processes (i.e. docking) for preliminary testing. As such the issue that arose from training on a conventional model, was relating to the extracted HADDOCK information. Information extracted through docking is somewhat redundant to the model's binary characterization, as information on contact points between the two molecules would already require docking and directly determine binding occurrence. This meant that, to employ docking information without performing docking, the singular model would have to be able to implicitly predict the 'point-of-contact' (AIF/RIF) features that resulted from docking data analysis. Conclusively, the singular model was trained with all features, except the AIF/RIF features that described points of contact between aptamer and target.

However, to further explore possible applications of these docking-related features, the idea of a stacked ensemble model was proposed. The stacked model would be divided in two segments, one responsible for predicting a set of points of contact between aptamer and the target (base models), and the other to translate those predictions into a binary classification (meta model). The selection of these points of contact was done by performing a preliminary training of a XGBoost model, using a complete dataset that contained all the AIF/RIF information. This allowed for analysis of features organized by relevance, which in turn allowed for selection of a set of AIF/RIF entries. The base models were then trained to individually predict all the 19 AIF/RIF features, while discarding the rest.

The next step consisted of exploring the two feature selection approaches with the selected model - Variance threshold (VT) and minimum redundancy maximum relevance (mRMR). This process was done experimentally before the optimization process. The first, tackles the dataset in a more statistical manner, narrowing it down in a way that fits a predetermined variance threshold. In this case, the threshold was set to 0.4, which means that all variables with variance lower than 0.4 were removed from the dataset. Minimum redundancy maximum relevance, as the name describes, attempts to select a number of features, while optimizing them to be the most descriptive and less redundant possible. These values were obtained experimentally through trial and error, and selected based on their impact to the model. The goal was to have a model that seemed optimistic with the selected dataset, and make it more conservative during optimization. Due to the lack of available data, models often tend to be overly optimistic within the training data, under-performing when exposed to unseen samples.

Optimization followed a similar principle to that of the preliminary model architecture testing. XGBoost has a wide range of modifiable parameters, and in this phase a set of 7 parameters were iteratively tested. The chosen parameters were: learning rate, sub-sample, column sample by tree, number of estimators, maximum tree depth, minimum child weight and gamma - chosen based on their expected impact on performance and learning. These parameters were tested with each being individually set to a certain range, the model trained and results compared. Promising or otherwise related parameter modifications were combined and again tested. This section was developed under the assumption that the model would tend to be over optimistic, as

the preliminary testing pointed towards it. As such, most optimizations were made with the goal of making the model as conservative as possible, while maintaining a good AUC, log loss and accuracy.

3.6. Feature analysis

Further model analysis encompassed observation of feature influence while training the model, specifically the newly added SurfMap features - as the point of contact features were removed from modeling. These were tested in the most promising model, selected based on the previous steps of this methodology, by removing them from the training dataset and observing the impact in model performance. With the goal of understanding which features better allow for correlation of aptamer target interactions, a process of feature analysis was conducted. This was done through the exploration of feature coefficients determined by the best performing models in the three dataset approaches: no feature selection, VT and mRMR.

3.7. Model Evaluation

Model performance was analyzed through the usage of traditional metrics present in the classification report module present in the sci-kit learn library: precision, recall, F1 score, accuracy and MAE. To complement analysis during training and testing, an AUC curve and log loss curve were plotted.

For final testing, a set of six models were stored (MA1, MA2, MB1, MB2, MC1, MC2). These models were stored based on their feature selection process (no feature selection, variance threshold and mRMR - A, B and C respectively), and the presence of SurfMap features during training (1 for present and 2 for absent). The final testing process relies on exposure of the models to an unseen dataset. This testing dataset was created and stored separately from the original, but following the same protocol.

3.8. Model implementation and application

With the goal of increasing accessibility and usability of this model, a web app interface was developed and published to Github. Developed in python, it takes four possible inputs: aptamer sequence, target sequence, target Uniprot id (optional) and mode of execution. The possible modes of execution are: target search, aptamer search and classification. The first takes an aptamer sequence as input, and cycles through the available protein targets classifying their interaction; the second does the opposite,

taking a target sequence or Uniprot id as input; the third takes both an aptamer sequence and target sequence (or Uniprot id), and classifies the occurrence of binding.

As an attempt to test model applicability in the context of marine bioremediation, AptaCom was used as a reference for possible E.coli targeting aptamers. A promising set of 5 aptamers were identified, all targeting whole cell E.coli, which means that the specific binding interaction is unknown. In the referenced work [40], researchers hypothesize that the aptamer most likely binds to a protein present in the outer membrane of E.coli specifically, independently of the strain. As an example of the model's application, the pair of proteins ompA and ompF were selected as potential targets to be tested against the referenced aptamers. The aptamer had its biochemical features extracted, following a pipeline identical to the dataset creation. The same will be applied to the protein targets. After extracting the required descriptors, the final model will be applied in the prediction of binding occurrence. The results will then be validated through HADDOCK docking, as a way of furthering model description.

Results

Following the structure of this work, results are divided into two sections, one focusing on the created database, its platform and preliminary information that can be extracted from them, and the other focusing the model, the process of its creation, its performance and feature information retrievable from it.

4.1. Database

The resulting database consists of approximately 4000 non redundant entries, describing both aptamers, targets and their interaction. Aptamers are described by their sequence information, chemical composition and publication name; targets are described by their publication name, sequence (when available), external id (when available) and by their chemical composition; lastly, aptamer-target interaction is described solely by their dissociation constant. Each entry directly relates to two source materials: the database from which they were retrieved, and the article from which they were originally compiled - described either by it's DOI or reference, according to available information. The final feature added relates to aptamer-target interaction validation, which in the context of this work was obtained through docking via HADDOCK.

In reference to the collected information, the following table describes some of the features present in the different databases. It's possible to observe that AptaDB stands out as the most complete in terms of both aptamer and target description. Despite the fact that NAKB relates back to the PDB database, it still lacks some aptamer specific information, such as binding affinity, or references on experimental conditions. Our database - AptaCom - employs the critical features that promote findability. The goal is to make entries as easy as possible to be scanned through and selected for further research.

As previously mentioned, all databases diverge in terms of approaches to data organization and description, which required different database schemes to be implemented in our final database. AptaCom employs the common features as searchable items, bridging the overlapping and diverging information across these databases. By employing these premises, AptaCom can better conform with FAIR principles, making information easier to find and access. It also allows interoperability between databases due to its commonality of features and references. And the workflow detailed in this work makes it more easily reproducible.

Features	UTexas	Aptabase	AptaDB	Aptamerbase	Apta Index	NAKB	AptaCom
Aptamer	X	X	X	X	X	X	X
A. Sequence	X	X	X	X	X	X	X
A. Chemistry			X		X		X
Target	X	X	X	X	X	X	X
T. Chemistry			X	X	X	X	X
T. Sequence			X				X
T. External ID			X			X	X
Binding Affinity	X	X	X	X	X		X
Ref/DOI	X	X	X	X	X	X	X

Table 1 - Feature comparison across databases

Of AptaCom's 4000 entries, there are a total of 2864 unique aptamer sequences and approximately 1805 unique target names. Of the targets, approximately 36% have an external ID that traces back to a database such as Uniprot, PubChem or ATCC. The 4000 entries are composed of entries from 6 different sources: UTexas aptamer database, AptaIndex, Aptabase, Aptamerbase, AptaDB and PDB. The largest sample belongs to UTexas with around 1400 entries, followed by AptaDB with approximately 1300 entries. AptaIndex and Aptamerbase both have approximately 730 entries. In the case of Aptamerbase, despite their almost 4000 entries, their entire database can be translated into only 725 aptamer-target interactions. Aptabase has nearly 400 entries, while NAKB has approximately 138.

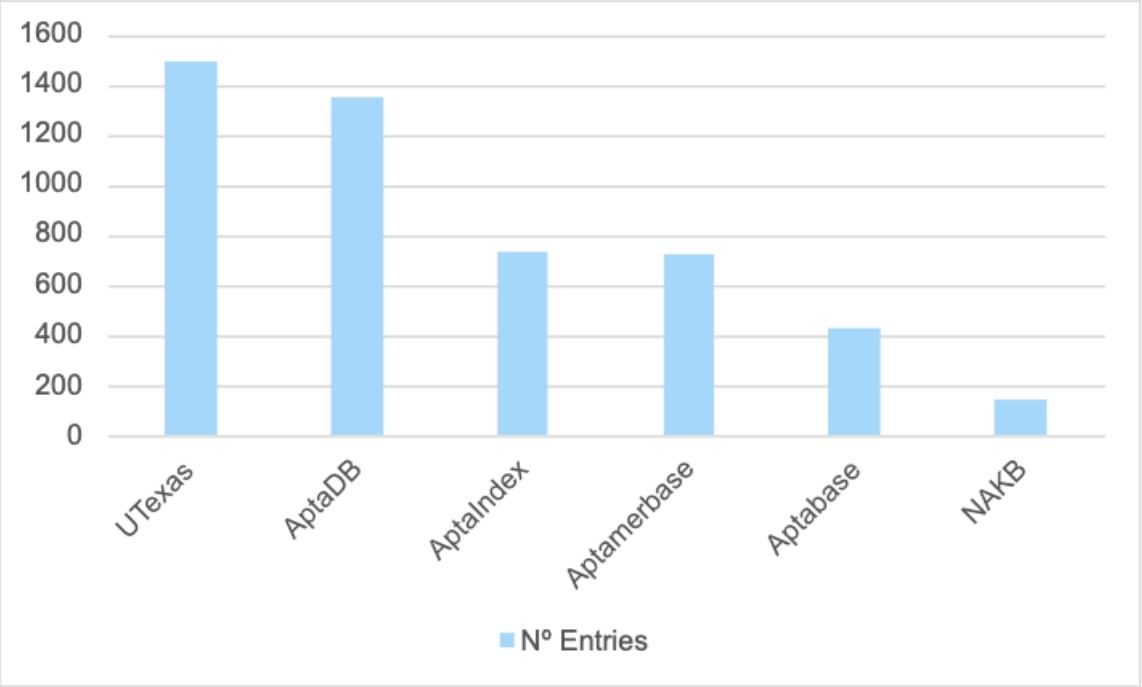

Figure 2 - Number of entries per database within AptaCom

Of the retrieved entries, a rough analysis was performed to understand the prevalence of target description in terms of AptaCom itself. The resulting chart describes the percentage of entries that have a well described target that traces back to an external database with further details. It's possible to observe that from the 4000 entries, only 38% have an external target ID - which represents approximately 1500 entries. The remaining entries describe the target only in terms of its name, often lacking even a specific chemistry description.

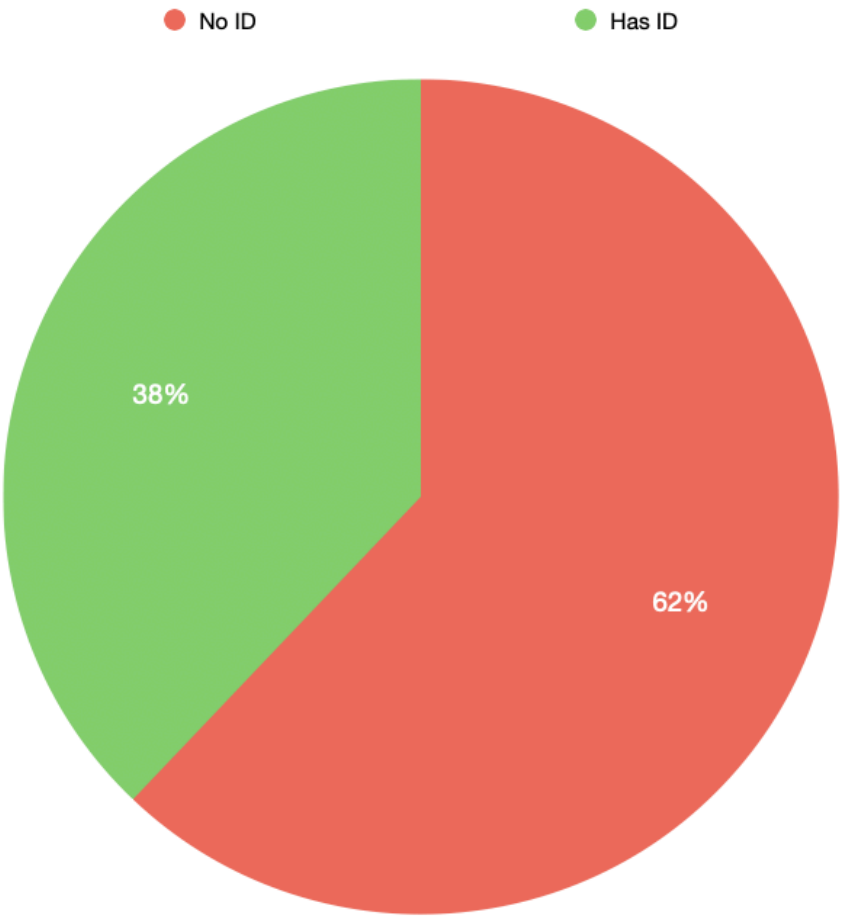

Figure 3 - Ratio of entries with external ID: 'No id' represents percentage of entries with absence of external ID that link to target specific database; 'Has ID' represents percentage of entries with such external ID.

Regarding target chemistry, it was observed that the great majority of targets have no discrete description of their chemical composition. The remaining categorized targets are distributed in the following order: protein, small molecules, cells, other, peptide, antibody, tissue, nucleic acid. As it can be observed, 46% of the entries lack target chemistry description - around 1840 entries - mostly from UTexas aptamer database. Of the remaining entries, 26% describe protein targets, 16% describe small molecules and 7% describe whole cell targets. Peptides represent only 1% of the described aptamer targets, while 'other' - which references AptaDB's classification of a target that does not fit in any of these categories - represents 4%.

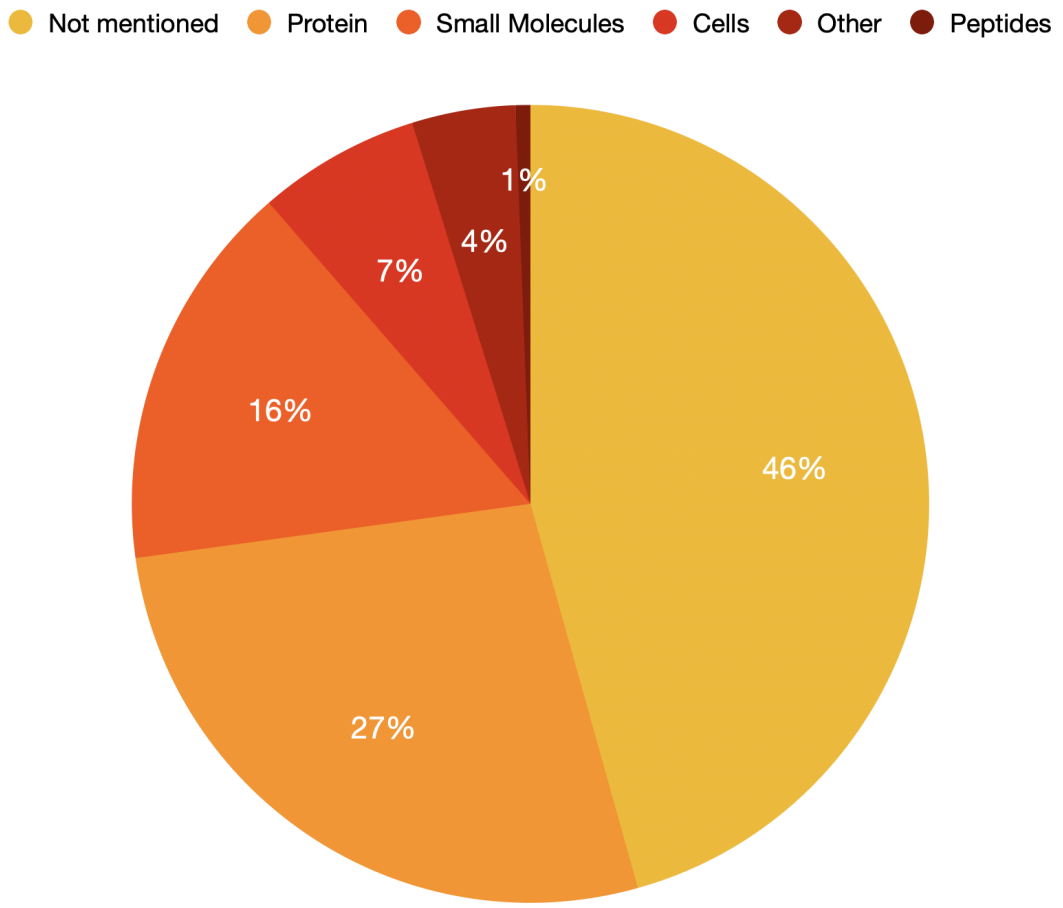

Figure 4 - Target chemistry distribution: 'Not mentioned' - there is no mention of target chemistry across databases; 'Protein' - percentage of entries with described as proteins; 'Small Molecules' - percentage of entries with described as small molecules; 'Cells' - percentage of entries with described as whole cells; 'Other' - percentage of entries with described as none of the others; 'Peptides' - percentage of entries with described as small peptides;

Due to their high malleability, aptamer chemistry can be very diverse. As a result, they are often categorized in the databases by the specific type of modification performed to the nucleotide. As a general representation, aptamer chemistry was grouped into DNA, RNA, Modified DNA and Modified RNA. In the following chart it is possible to observe that 52% - representing the largest portion of the database - describes DNA aptamers. Approximately a quarter of the database consists of aptamers with non-described chemistry. Due to the simplified way aptamer sequences are usually represented, it is difficult to manually categorize these entries without further information on the publication. The remaining entries describe RNA aptamers - 20% - and modified RNA aptamers - 5%.

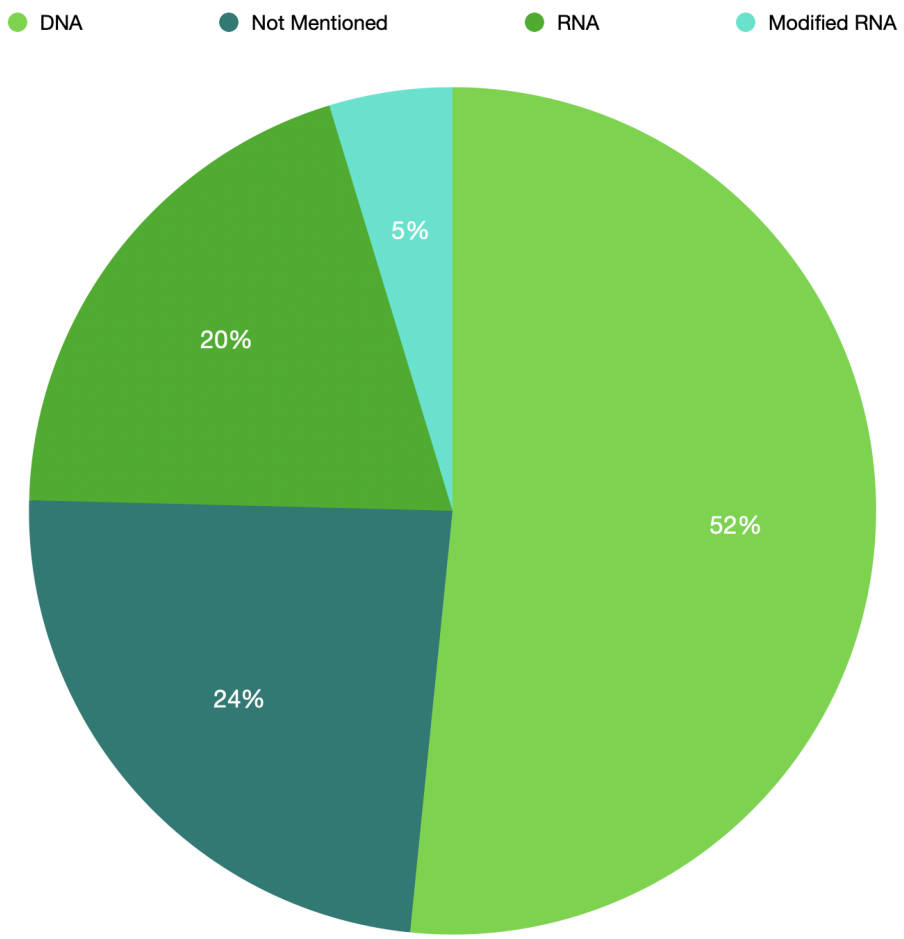

Figure 5 - Aptamer chemistry distribution: **'Not Mentioned' - percentage of entries without chemical description; 'DNA' - percentage of entries with described as DNA aptamers; 'RNA' - percentage of entries with described as RNA aptamers; 'Modified RNA' - percentage of entries with described as modified RNA aptamers;**

Despite most databases not mentioning experimental or computational validation of their entries, as a part of the modeling process, there was a docking phase, which in turn allows for some degree of preliminary validation of entries. This in turn permitted preliminary validation of entries, which can later be expanded upon, by following the same process. Of the total 4000 entries, approximately 16% were computationally validated via HADDOCK docking, which represents all aptamer-target pairs tested via HADDOCK.

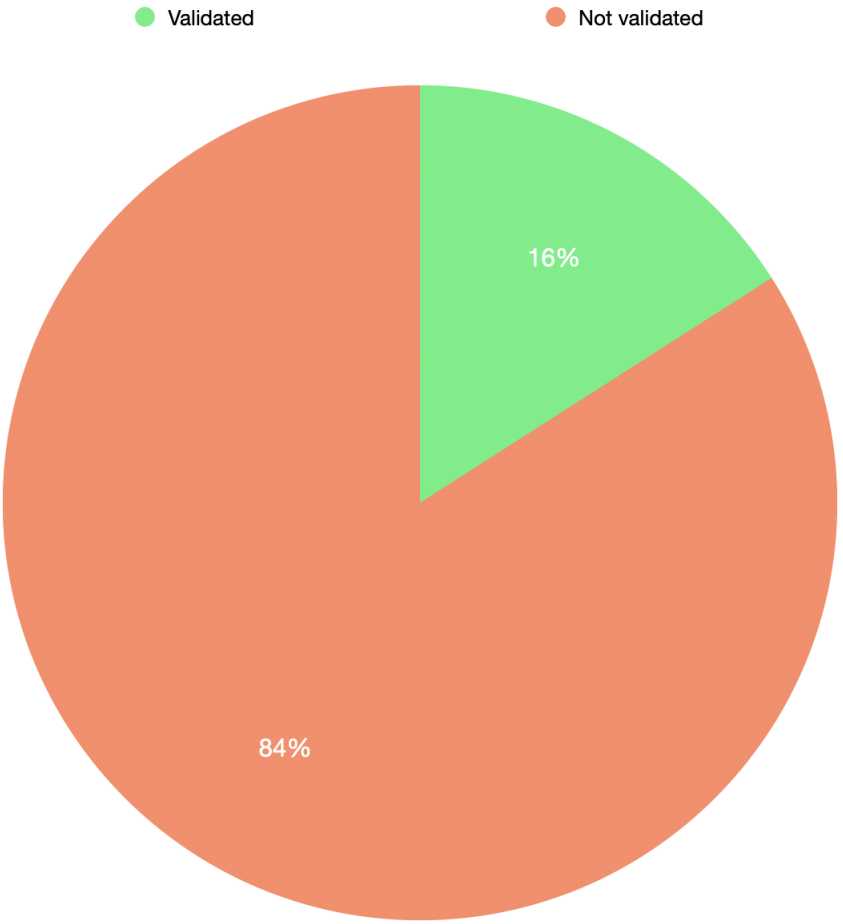

Figure 6 - Portion of database validated through docking: 'Validated' - percentage of database validated through docking; 'Not validated' - percentage of database not validated through docking;

4.2. Dataset

The dataset was constructed by extracting all features made available by the PyBioMed library, regarding both sequence structure information and physico-chemical characteristics; along with surface features obtained through SurfMap and AIF/RIF features extracted through analysis of HADDOCK results. The resulting data set consisted of a total of approximately 12500 columns, comprising 996 entries that described both positive and negative examples. The 'compressed' information (300 columns) was added, resulting in a dataset with approximately 12800 columns

4.3. Model

The first stage of preliminary modeling saw evaluation of different model architectures with the dataset, resulting in 4 metrics, previewing model performance. In this context, the models explore a dataset without AIF/RIF features; the models in question employ their default parametrization. From a baseline model, the best performing were the ensemble models, with XGB slightly outperforming RF with 0.870 and 0.850 of accuracy, respectively. SVM achieved a promising 0.800 of preliminary performance, while ANN achieved 0.670.

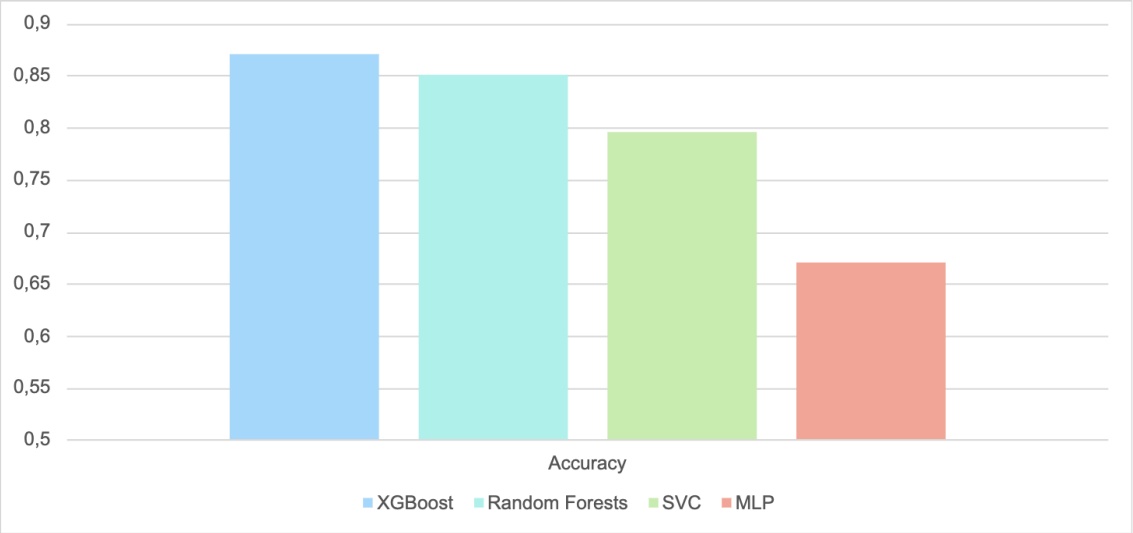

Figure 7 - Preliminary model overview: Describes the results obtained by each model with standard parameters and a non-treated dataset;

Exploration of a stacked ensemble architecture can be broken down into results from the base models and from the meta model. The 19 base models displayed uncertain performances, in relation to the target, however the large majority barely surpassed chance. The meta model, pointed to over-fitness, reaching accuracy of 100% without any type of optimization. An overview of the two best and two worst performant base models (B1- B4), as well as the resulting meta model metric (B5) can be observed in the chart below. The almost perfect performance of the meta model is a good indicative of over-fitting, especially in the context of the performance of the other base-models.

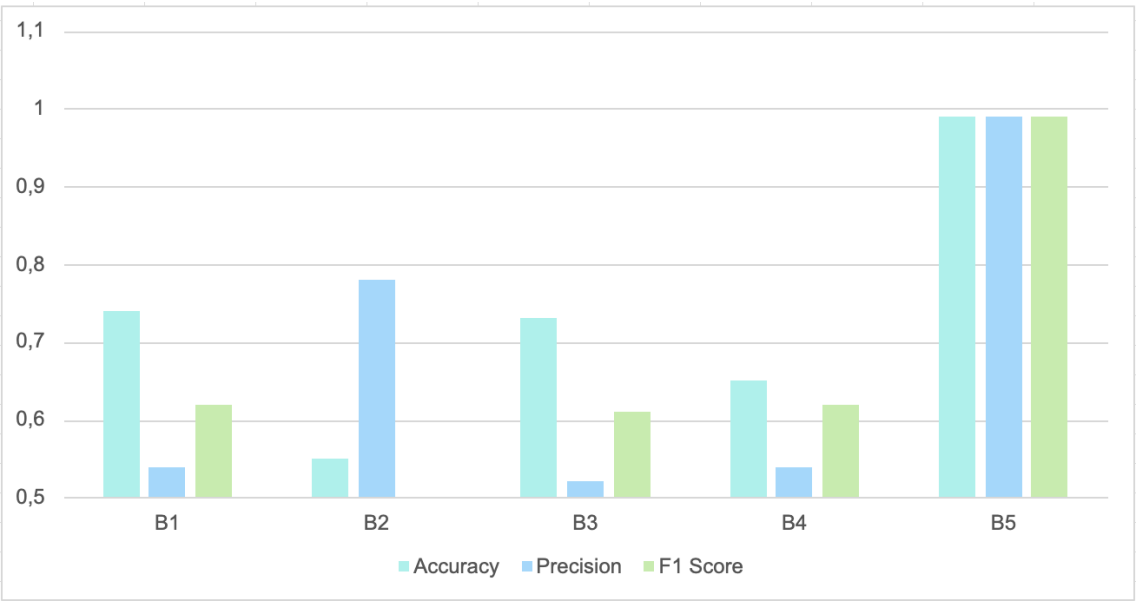

Figure 8 - Metrics of four (two best and two worst performing) of the nineteen base models (B1-B4) that served as basis for the meta model (B5). These models were described in terms of their accuracy, precision and F1 Score.

4.4. Optimization

Feature selection was performed, rendering a set of three models: MA1 (no feature selection), MB1 (feature selection through variance threshold), and MC1 (feature selection through minimum redundancy maximum relevance). These models were trained and stored for posterior bench-marking. The three models were analyzed in terms of the accuracy, F1 Score and AUC in both training and testing datasets. Here, it was possible to observe the discrepancy between model performance in the training and the testing datasets. MA1 showed the highest accuracy discrepancies between the

training and testing datasets, with accuracy dropping approximately 13% (from 0.875 to 0.744). The remaining models saw less discrepancy with MB1 performing slightly worse at 1% (0.885 to 0.875), similar to MC1 (0.792 to 0.771). If MA1 is used as a reference, MC1 is the one that showed the highest performance drop during training, with MB1 outperforming both MA1 and MC1 in all metrics.

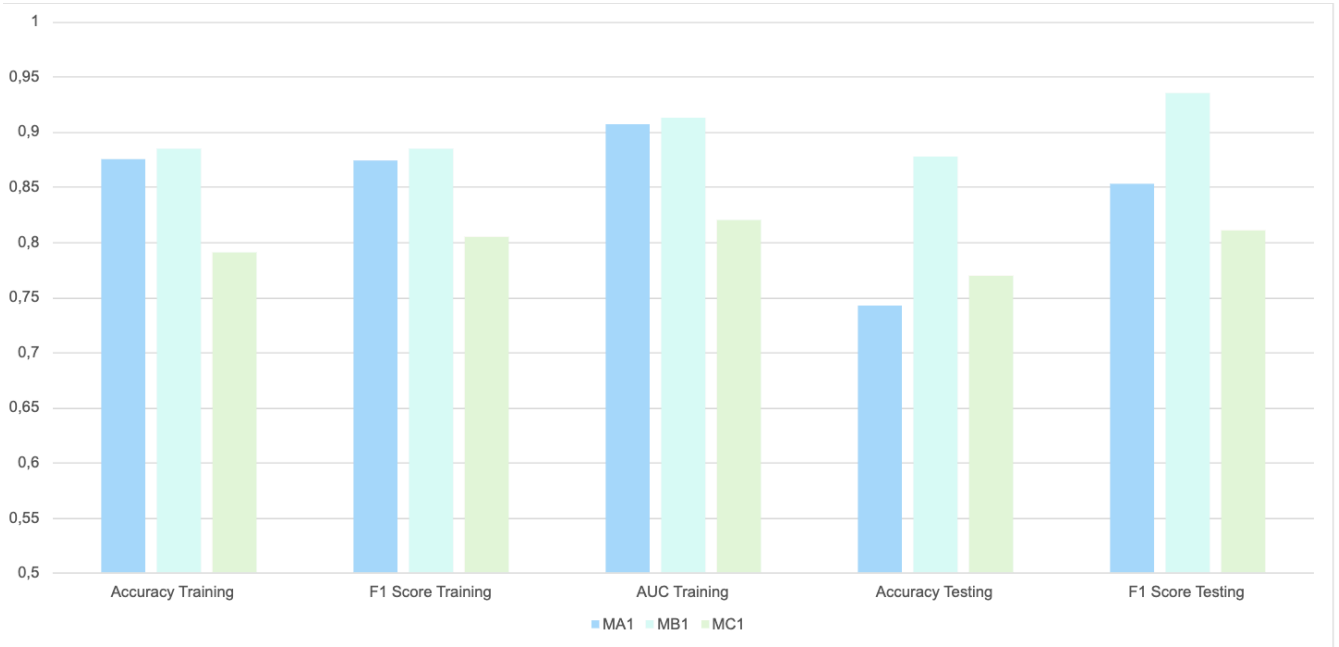

Figure 9 - Evaluation of feature selection approaches: MA1 - no feature selection; MB1 - variance threshold selection; MC1 - minimum redundancy maximum efficiency selection.

The process of optimization rendered an extensive array of different performances. Modifications were made with the goal of reducing expected variance. In figure 10, it is possible to observe the various results obtained in six of the assays performed. The first (M1) describes the default XGB performance, while the second (M2) explores the modification of the sample and the number of descriptors. From then on (M3-M6), the goal was mainly to restrict the model. So, the parameters that improved the model variability were balanced with parameters that would make the model more conservative. The Area Under the Curve (AUC) and Loss metrics were used as a way to track drastic impacts in performance. As observable in the graph below, both remained relatively stable after the addition of more restrictive parameters. The last model was the best performative of the restricted batch, with an accuracy of 0.885, F1 score of 0.874, AUC of 0.909 and log loss of 0.125.

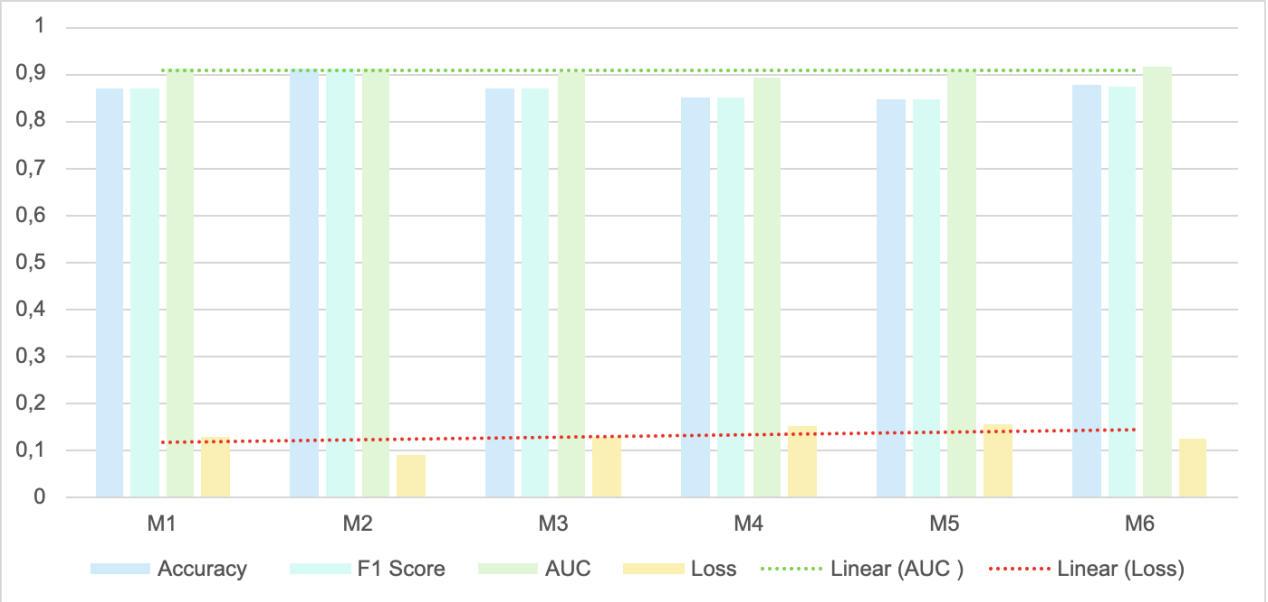

Figure 10 - Describes 6 assays performed, in order to make the XGB model more conservative. Assays M1 and M2, represent a default and more adaptive models, respectively. M3 to M6 describe models with an increase in restrictive parameters.

Following the performance evaluation in relation to the feature selection process, the impact of SurfMap features was analyzed, in the previously best performing model (MB1 versus MB2). As observed in the chart below, despite slightly outperforming MB1 in the training dataset, the discrepancy between training and testing metrics is much higher, with MB2 dropping approximately 10% in accuracy. There were no significant changes in AUC and F1 Score in between datasets in regards to MB2. However, MB1 did show a significant drop in F1 Score from training to testing of approximately 4%.

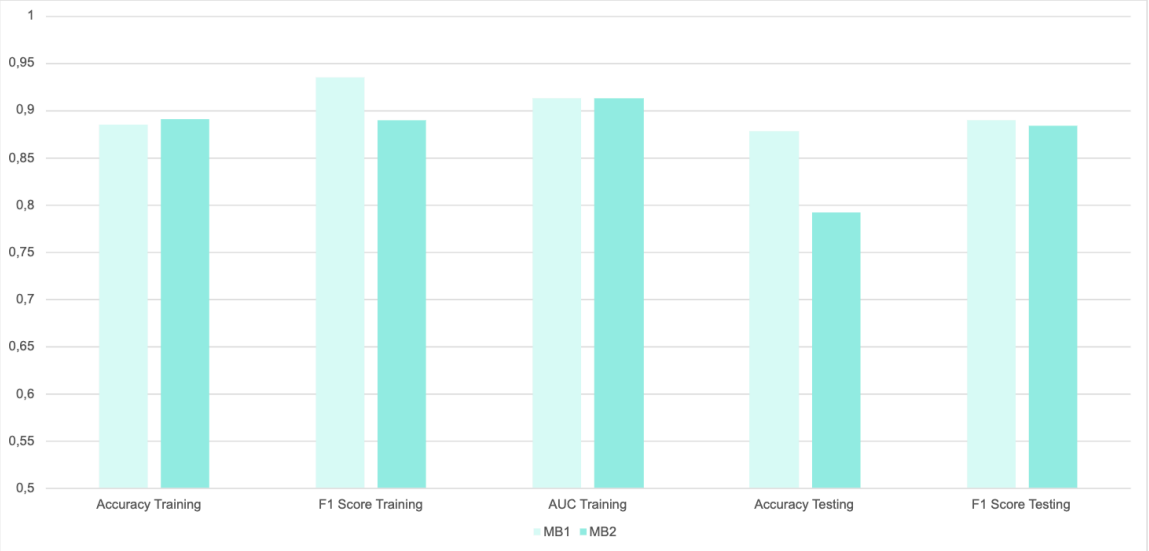

Figure 11 - Evaluation of influence of SrufMap features in model performance: MB1 - model trained on SurfMap features; MB2 - model trained without SurfMap features.

4.5. Feature relevance

The following describes an overview of the feature distribution of the feature selected by the XGB model during training. This analysis was done on the best performing model. SurfMap features were the most abundant in the dataset, as well as the highest scoring features in the model. The protein sequence features (P.S.Features) describe an array of physico chemical characteristics, from polarity to hydrophobicity to charge, taking up to 32% of the selected features. The lowest scoring, and least abundant features related back to aptamer kmers, taking only around 1% of the selected features.

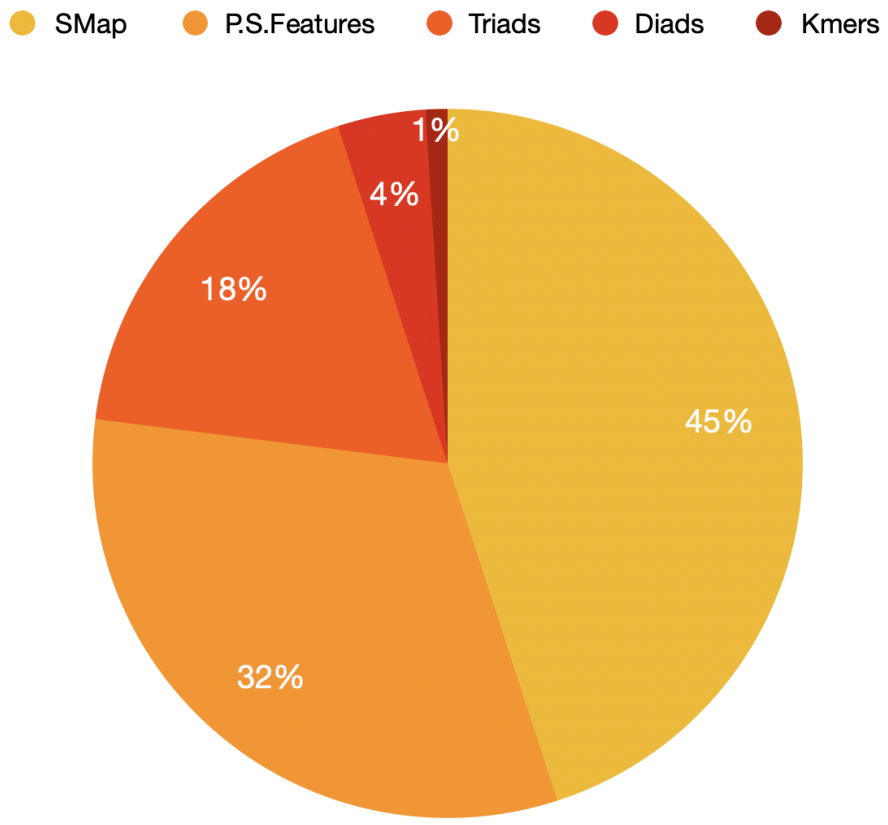

Figure 12 - Most relevant features during training: 'Smap' - SurfMap features; 'P.S. Features' - physio-chemical features from protein sequence; 'Triads' and 'Diads' - protein sequence tokens of three and two respectively; 'Kmers' - aptamer sequence tokens of sizes two and three;

4.6. Model Application

The five selected aptamers paired with the two targets resulted in 10 possible binding/non-binding scenarios. After model implementation, it was possible to observe that of the 10 samples, 9 were identified with a high probability of binding to the targets. When compared with the binding results obtained through HADDOCK, it was observed that 8 out of the 10 samples performed binding.

Method	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
HADDOCK	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Model	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

Table 2 - Model application and HADDOCK validation: (P1-P10) - assays performed between five aptamer and ompF (1-5) and ompA (6-10); HADDOCK - prediction of binding based on ZDOCK docking score; Model - prediction of binding based on binary model classification;

Discussion

This work was developed with two broad goals in mind: development of a database that compiled different sources of aptamer information into one singular hub, and development of a machine learning model capable of predicting aptamer-target interaction.

5.1 Database

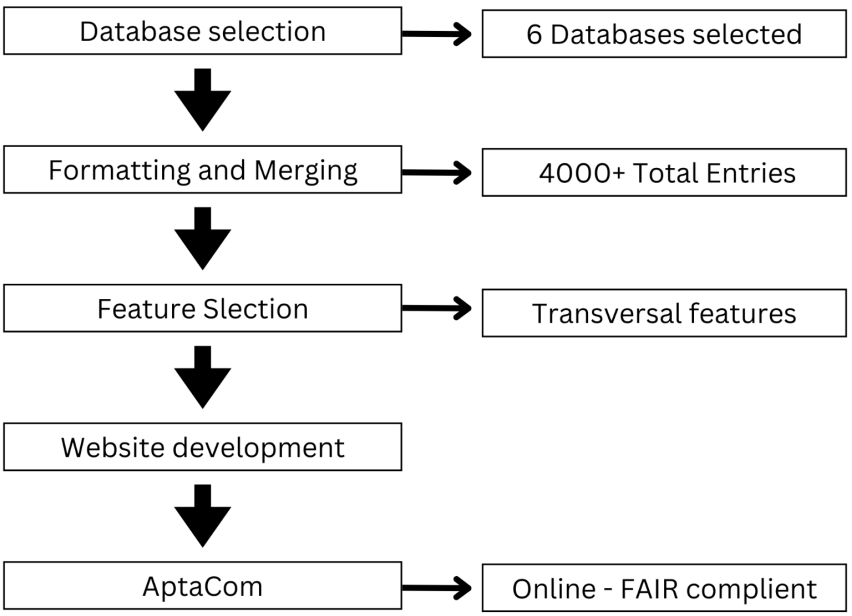

Figure 13 - Summary of results obtained throughout database development

The current landscape of aptamer databases can be described as incredibly diverse, and dispersed. Despite various available sources of applicable experimental information, databases are created under different premises, by different groups with no standardized oversight. This data dispersal, and its associated lack of standardization, raises difficulties to preliminary research and information availability, this in turn can affect accessibility to information for both researchers of the field and new-comers alike, potentially dampening interest in the area. With the goal of combating these issues, by bridging together these different data sources, AptaCom was developed. Its premise is to compile and combine data from an extensive array of sources, into one singular, non redundant database, in turn making their data more

easily accessible and comparable across sources. The resulting AptaCom database spans 4000 entries, with 16 features describing its components. This data compilation allows for a broader overview on the aptamer research landscape, while also conforming with FAIR principles, as it is standardized, downloadable and comparable in its entirety. AptaCom is composed of the combination of 6 different databases: UTexas Aptamer database, Aptamerbase, Aptabase, AptaIndex, NAKB and AptaDB. Aptabase is the simplest database of the set; its entries describe target and aptamer name, along with binding affinity - often with often missing or unintelligible units. UTexa's database emphasizes in promoting well described experiments, specifically sequence modifications applied to the aptamer throughout its development. Even though there is a detailed description of the aptamer sequence and its pool type, there is no description of the target other than its1 name. AptaIndex presents itself as a sort of an aptamer catalog, however it describes important features, even including aptamer molecular weight, secondary structure image and binding conditions. The key factor missing remains a more detailed description of the target. NAKB serves primarily as a sort of index for information gathered in the PDB website, but with emphasis in nucleotides. As such, it follows a similar structure and feature set with that of PDB, with targets being linked to either a three-dimensional PDB file, or even an Uniprot id. Despite the completeness of its data entries, the 'affinity' feature is rarely described, which in turn would require either computational simulation or experimental testing to retrieve. Aptamer base is a very specific case, as it was hosted on a platform that followed a relational structure. That database is now offline, and its relational data is no longer available. Regarding the aptamer database itself, the available files describe aptamer and target interaction similarly to what was done by AptaIndex and UTexas, with light description of aptamer sequence and target name along with their binding affinity. A feature that displays promise, specially in combating the reproducibility issues of the field is the 'validated by', which relates to other papers that also tested the aptamer-target pair in the entry. Finally, AptaDB is by far the most complete database described in this work. It organizes data with a relational approach of isolated databases, connected by a respective key. It describes both aptamer and target in detail, in terms of their chemistry, sequence and binding affinity between the two across various well described assays. The aim of AptaCom was mainly to improve information accessibility, with focus on making aptamer information more easily searchable. Considering the most frequent features across databases, along with other important

components while researching and understanding aptamer-target interactions, the chosen features were: aptamer and target name; aptamer and target sequence - when target sequence (i.e. smiles or amino acid sequence) are available; binding affinity in nM and a link to the original database, along with either reference or DOI of the original publication. Additionally, due to the differences across databases, two other features were added: key-word matching - which tokenizes aptamer and target name and makes their tokens searchable - and sequence similarity - done in AptaDB to allow for comparison of similar aptamer sequences. The AptaCom database was built not to replace currently available databases, but to bridge their knowledge into one source. By combining the data into one accessible platform, there is also opportunity for a broad spectrum analysis of the aptamer research landscape, especially regarding the issue of standardization. Across the six databases, only around 36% display a specific external id that describes the target in detail. This is to be expected due to the nature of aptamer research, and its emphasis on applicability more than how the molecular process develops. Despite this problem being unsolvable without an extensive manual review of aptamer publications in search of a detailed target description that allows tracing back their id, sequence or structure, AptaCom highlights when such key information is missing, hopefully setting a standard on how to publish aptamer information- whether it be articles or database entries. Another component neglected by some databases is ’Target chemistry’, of which more than 2000 entries lack. These entries belong mainly to UTexas, and Aptamer base, both databases center around the aptamer and the published research, neglecting the target.

5.2 Model

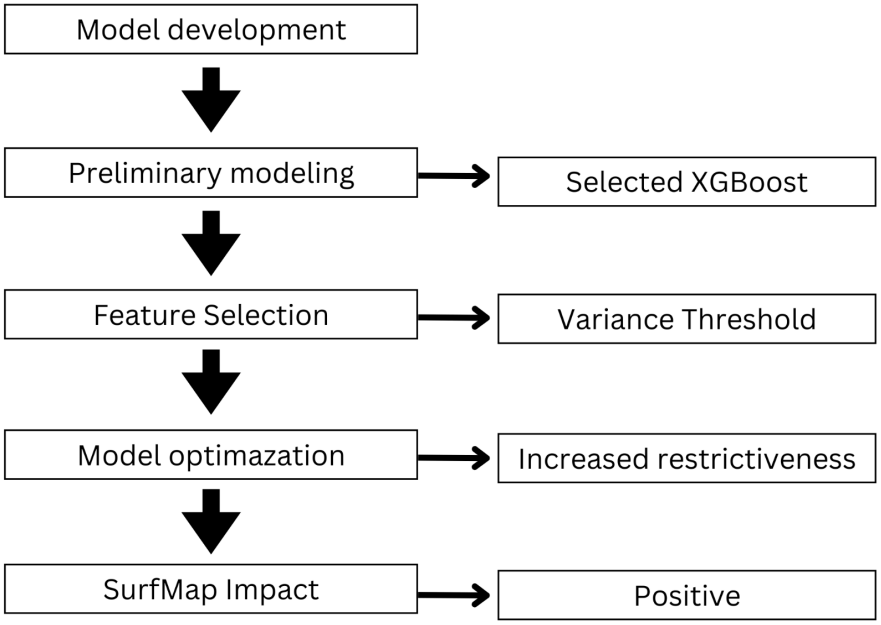

Figure 14 - Summary of results obtained throughout model development

The model was developed mainly with the goal of bridging the aptamer research accessibility gap by making preliminary research stages more accessible, allowing researchers to explore hypothesis without the requirements of a wet-lab. In parallel, the goal was to also further the understanding of aptamer-target interactions, through exploration of the influence of certain features in model behavior. The latter was done with emphasis on three dimensional and three-dimensional adjacent features, obtained through the usage of the SurfMap tool - describes protein surface in 2D space - and through a process of docking of aptamer-targets, and mapping of their contact points which lead to the creation of the metrics AIF and RIF - absolute and relative interaction frequencies, respectively. The process of model development can be segmented in two main sections: preliminary and optimization.

The preliminary stage was approached experimentally, with different model architectures being tested with the same extensive dataset, in order to choose a promising architecture. The first stage of preliminary modeling focused on analysing different model architectures, and showed that the ensembles behaved best, versus

ANN or SVM. This aligns with the literature, as the first aptamer models were based of an RF model. Additionally, ensembles tend to behave well with non-scaled datasets with lower sample size, while being are also very adaptive. Despite the flexibility of ANN's, these models require extensive datasets to perform optimally, which explains it's lower score. As for the SVM models, these tend to perform well in the conditions presented by this dataset: larger number of features while also being more features than samples. However, SVM's often under perform in datasets with a lot of noise or overlapping features between the classes, which is the case in this dataset. Possibly, further tailoring the features could lead to an improvement in performance of SVM. The ensemble models, RF and XGB, performed within the same range, with the XGB model outperforming RF in most metrics. It makes sense that these models would perform within the same range, but XGB's approach to inner feature relevance, as well as it's default parameters seemed to better align with the details of this dataset. Another advantage of this model is its extensive array of parameters, which allow further optimization and testing. As such, the XGB model was chosen for the next stages of development. During experimentation with different model architectures however, models often tended to display over optimistic results without requiring any optimization. This prompted further exploration through feature analysis of the models, in this case both RF and XGB were analyzed, as both allow access to a coefficient of feature relevance. In both cases, features regarding AIF/RIF showed high coefficients of feature importance, and the array of selected features tended to show little diversity. This was to be expected, as these features are redundant to the models prediction occurrence of binding between aptamer and target, thus being removed from further experiments with this model structure.

Information describing points of contact between an aptamer and a target could make not only an interesting feature for aptamer-target classification, but also for an independent model itself. This triggered the exploration of the stacked ensemble model structure, which employed a set of base models that fed into a meta model. The base models targeted the 19 most relevant points of contact - of 104 total. Their performances were scattered, with the majority barely surpassing chance. However, the meta model is still over-fitted. These results again, were probably due to various components, mainly the small sample size. These three-dimensional features represent exclusively positional information in three-dimensional space. In the context

of complex molecular interactions, it's probable that this sample size was not able to represent significant amounts of patterns that allowed deeper connections between the features in the feature set and the points of contact. This premise and features are still interesting for future developments, even if not applied in this model structure. This feature can in theory provide an overview of aptamer-target interaction, which could be useful in wet-lab or modeling optimization.

The process of feature selection is often key for better tailoring a machine learning model. As observed in the results, the application of VT feature selection resulted in the best performing model with the training and benchmark datasets (0.885% and 0.875% respectively). This is thought to be due the size of the feature set, and a proportional amount of noise. The removal of all features with variance below 0.40 resulted in the removal of 93% of features, with only 864 features remaining. This lack of variance across the dataset, can be due to two main factors: sample size and excess noise. The other approach taken was mRMR, which is an algorithm expected to lower redundancy while also maximizing the information carried by the features present in the dataset. However, when tested with a number of features similar to that of the VT approach (800), it negatively impacted model performance. Traditionally this method has issues dealing with higher dimensionality data, as well as non linearly correlated data, which justifies this loss in performance. Not performing any kind of feature selection also impacted the model negatively. The model performed more optimistically with the training data than with the benchmark (0.860% versus 0.700% respectively). The discrepancy between performance in training and with benchmark dataset is again probably due to noise in the dataset. The application of a 'blind' feature selection (such as VT) based solely on statistical characteristics of the data, combined with the feature scoring of the XGB, allowed the model to prioritize more information packed features, resulting in the better performing model (MB1).

The best performing model (MB1) was a XGBoost model, with the dataset subjected to feature selection through VT (threshold of 0.4). All the preliminary analysis of the models, pointed to the probability of an over optimistic performance with the training dataset. As such, the process of optimization was done with restrictiveness in mind. The resulting model surpassed Li. et al's performance in their 2014 model [15] - 81.34% accuracy and 77.41% with the benchmark dataset - while following a similar data structure and while also using an ensemble model. Compared with AptaNet's [17]

performance of 99.79% accuracy with the training dataset, and 91.38% accuracy with the testing dataset; the final model's performance is lower. It's important to consider that the datasets vary wildly in size, with AptaNet's dataset being three times as large. Their approach shows promise in the application of a largely synthetic dataset, raising the question of possible improvements in performance of MB1 if trained on a larger synthetic dataset.

Regarding the added three-dimensional features, namely AIF/RIF and SurfMap features, these impacted performance very differently. As previously mentioned, AIF/RIF features lead to over-fit models. Without testing with a larger sample size, with more diverse interactions, it is difficult to say if this metric was negatively influenced by the sample size, or if the metric itself is non applicable in this context. The stacked ensemble approach allowed a preview of the correlation between contact features and the remaining dataset, but did not prove successful in breaking the meta model's over-fitness. The premise of analyzing and translating some degree of molecular three-dimensionality remains relevant and should be explored in future works. SurfMap features showed great relevance in all models, as it represented not only the majority of selected features - independently of feature selection approach - but those with highest coefficients. Additionally, it was possible to observe that, despite performance remaining similar with the removal of SurfMap features, when testing with the benchmark dataset, it drops drastically, which points to the positive influence of these features in the dataset. Surface characteristics of proteins play a big role in shaping its molecular interactions, as previously mentioned a good portion of aptamer-protein interactions is shaped by surface charge and shape [4].

Regarding the percentage taken by aptamer features in the best performing model (approximately 1%), it's important to consider that those features were selected both by the VT process and by the XGBoost model. It was expected that aptamer features would take a smaller portion of the feature set, as their numbers are much smaller than the number of protein features. However, its position as the smallest portion raises critical questions on their part in model prediction. The array of features regarding DNA and RNA sequences provided by current modules, don't seem to be as wide reaching as protein features, which leads to a disparity that can be identified during modeling. Expanding upon aptamer characterization specifically for modeling could improve future model performances.

The application of the model in the context of a sample bioremediation example aimed to employ the tools developed throughout this work - database, feature extractor pipeline and model. This was done with the goal of exemplifying this work's application in the context of marine bioremediation, by targeting a common health hazard - E.coli.

The five selected aptamers targeted E.coli as a whole cell - whole cell SELEX technique - which doesn't indicate a specific protein target. Based on the premise that the aptamer would bind to an exposed protein belonging to the outer membrane [40], the ompA and ompF proteins were selected, and tested using both the model and HADDOCK - the latter as a validation technique. The model was capable of correctly identifying 9 of the 10 cases in terms of binding/non-binding interactions.

This performance followed expectations, including the difficulty in identifying correctly negative samples. Aptamer models tend to be too optimistic in their predictions, often performing worse with negative samples than with positive [18]. Besides its relation with data balance, this performance issue can be caused by model design or even feature selection and relevance. In the context of this work, the proportionally smaller number of features describing aptamers when compared with features describing proteins can point to a less efficient description of the aptamers. This broader aptamer description can mean that some critical features may not have been properly described, affecting model behavior. To combat this it would be viable to expand the dataset, and possibly guarantee that a certain percentage of the selected features corresponds to the aptamer, either by filtering out target features or by expanding nucleotide descriptors.

One of the questions raised in the beginning of this work, regarded the choice, or lack thereof, of datasets in the development of previous machine learning models. Ideally, the analysis of the impact of the AptamerBase dataset versus the dataset employed in this work, would require cross-testing the models with the respective datasets. However, due to time constrictions and availability of some of the models in question, such analysis was not possible to perform. This remains an interesting analysis for future work, as the field of aptamer research progresses.

Conclusion

This work was developed with one broad spectrum goal in mind: to improve accessibility of aptamer research. This goal was divided into two components: a database that compiles and organizes information dispersed across different sources, and a model that could act as an auxiliary tool for preliminary research.

6.1 Database

AptaCom (https://jc-biotechaiteam.com/AptaCom/) is an open source database created with the goal of integrating various data sources, allowing researchers to leverage different approaches to research, while also creating a space for standardization. It allows for preliminary research based on: aptamer sequence, target sequence, target external ID, reference, DOI and database of origin. As of now, the AptaCom platform is in its early stages, but the main future goal is interoperability and automation. The work put into creating each of the source databases was extensive, with reviewers analyzing thousands of papers combined. This was due to the lack of a singular platform and norm of adding such information to a database. To combat that, AptaCom contains a form page that will be key for standardizing, reducing friction of the publishing process and hopefully promoting more stable information dissemination. In an attempt to prevent loss of relevance from the database one of the future goals is to add an automated information retrieving process. Using the already existing pipeline for retrieving information from currently existing active databases, the goal would be to bi-annually retrieve information and add it to AptaCom automatically. This would allow the continuous distribution of the work carried by these institutions, while also maintaining the database up to date and relevant.

From the data collected in the database, it is possible to observe the broad reach of aptamer research, and the relevance of this work in bridging information accessibility for future researchers.

6.2 Future Research

Besides its more direct application as a source of information for research, a database of this scale can serve as a basis for future research. It opens doors for research that tackles, for example, aptamer sequence composition versus target chemistry, or alternatively an analysis of secondary structure and binding affinity across classes of targets. In addition, it also allows a broad overview of the landscape of aptamer research, in which biases towards specific target types or aptamer compositions could highlight blind spots in the research worthy of further exploration. These opportunities arise from merging information collected across decades of research in one single platform.

6.3 Model

The machine learning model developed in the context of this work, performed within expectations. Its development allowed a deeper analysis of the positive impact of surface features in model performance, while also prompting a review of other physio-chemical features relevant for aptamer-target interaction. It was possible to observe that the addition of a metric that describes aptamer target contact in specific points may be very useful for a sufficiently big dataset, despite causing over-fitting in the dataset at hand. The addition of surface descriptors did prove useful in improving an otherwise average model, which could point to a future model where a larger array of model descriptors relate back to three-dimensional features. The process of feature selection proved useful in trimming excessive noise present in such an extensive dataset.

6.4 Future Research

Due to time constrictions, it was not possible to verify the influence of the methodology employed by Li et al. and subsequent works. As such, future works should focus on analyzing their pipeline and dataset and if validated, merge it with the dataset developed in this work, to expand on future modeling.

There is still much room to grow in terms of in-silico modeling of aptamer-target interactions. This is especially true in terms of possible descriptors and other modeling architectures. The relevance of features relating back to target surface information points to a possible key role that three-dimensional features describing the aptamers could have. Additionally, in a similar way to how three-dimensional features influenced

the model, it's possible that the secondary structure (of both target and aptamer) could improve model performance, adding another layer to interaction description. Stemming from the RIF metric, another possibly useful approach to the description of three-dimensional structure of aptamer and target, would be something along the lines of Relative Exposed Residue Frequency (RERF). Such a metric could be used to describe to the model the rate of exposed nucleotides and of amino acids, based on their PDB file. It would require the creation of another tool or pipeline, but could prove useful, especially when paired with something like the SurfMap tool. Alternatively, simpler options such as better tailoring the physio-chemical features already obtained, through application of other feature selection processes.

It was clear that modeling in-silico aptamers can prove useful for accelerating a pipeline for bioremediation development. Furthering this research would require secondary experimental work to validate computationally obtained results in regards to the specific target interaction.

The aptamer analyzed in this work, was already studied in the context of biosensor development [40]. This validates the premise of application of the infrastructure built throughout this work as useful tools for aptamer development.

6.5 Final thoughts

This work focused on building upon existing knowledge of the field of aptamer research. It brought new tools and platforms that will hopefully serve as a stepping-stone for future researchers. The AptaCom database bridges an array of diverse sources which can help bring aptamer research closer. While the model resulted in promising conclusions of aptamer-target interactions, and can be used as a possible tool for aptamer development. In the context of this work, we focused on model application in the context of bioremediation. But these biomaterials are versatile and can be applied in a wide range of areas of research. As any other biomaterial, aptamers are limited by their definition and by the creativity of their developers and researchers.

Bibliography

[1] A. D. Ellington and J. W. Szostak, “In vitro selection of RNA molecules that bind specific ligands,” Nature, vol. 346, no. 6287, pp. 818–822, Aug. 1990. [Online]. Available: https://www.nature.com/articles/346818a0 [Cited on page 5.]

[2] C. Tuerk and L. Gold, “Systematic evolution of ligands by exponential enrichment: rna ligands to bacteriophage t4 dna polymerase,” Science, vol. 249, no. 4968, pp. 505–510, Aug. 1990. [Online]. Available: https://www.science.org/doi/10.1126/science.2200121 [Cited on page 6.]

[3] M. Darmostuk, S. Rimpelova, H. Gbelcova, and T. Ruml, “Current approaches in SELEX: An update to aptamer selection technology,” Biotechnology Advances, vol. 33, no. 6, Part 2, pp. 1141–1161, Nov. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0734975015000336 [Cited on page 6.]

[4] T. Adachi and Y. Nakamura, “Aptamers: A Review of Their Chemical Properties and Modifications for Therapeutic Application,” Molecules, vol. 24, no. 23, p. 4229, Nov. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC6930564/ [Cited on pages 7 and 48.]

[5] J. F. Lee, J. R. Hesselberth, L. A. Meyers, and A. D. Ellington, “Aptamer database,” Nucleic Acids Research, vol. 32, no. Database issue, pp. D95–100, Jan. 2004. [Cited on pages 7 and 10.]

[6] J. Cruz-Toledo, M. McKeague, X. Zhang, A. Giamberardino, E. McConnell, T. Francis, M. C. DeRosa, and M. Dumontier, “Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX experiments,” Database: The Journal of Biological Databases and Curation, vol. 2012, p. bas006, 2012. [Cited on pages 7, 9, and 12.]

[7] L. Chen, Z. Yu, Z. Wu, M. Zhou, Y. Wang, X. Yu, W. Li, G. Liu, and Y. Tang, “AptaDB: a comprehensive database integrating aptamer-target interactions,” RNA (New York, N.Y.), vol. 30, no. 3, pp. 189–199, Feb. 2024. [Cited on pages 7 and 10.]

[8] A. Askari, S. Kota, H. Ferrell, S. Swamy, K. Goodman, C. Okoro, I. Spruell Crenshaw, D. Hernandez, T. Oliphant, A. Badrayani, A. Ellington, and G. Stovall, “UTexas Aptamer Database: the collection and long-term preservation of aptamer sequence information,” Nucleic Acids Research, vol. 52, no. D1, pp. D351–D359, Jan. 2024. [Online]. Available: https://academic.oup.com/nar/article/52/D1/D351/7334094 [Cited on pages 7 and 10.]

[9] Vinay Bachu, Lazmy Deware, Ayush Kumar, Pooja Rani Kuri, Malaya Mili, Naveen Kumar Singh, and Pranab Goswami, “Aptabase: An aptamer database,” Oct. 2021. [Online]. Available: www.iitg.ac.in/proj/aptabase [Cited on pages 7 and 10.]

[10] A. A. Miller, A. S. Rao, S. R. Nelakanti, C. Kujalowicz, T. Shi, T. Rodriguez, A. D. Ellington, and G. M. Stovall, “Systematic Review of Aptamer Sequence Reporting in the Literature Reveals Widespread Unexplained Sequence Alterations,” Analytical Chemistry, vol. 94, no. 22, pp. 7731–7737, Jun. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9179646/ [Cited on pages 7, 8, and 10.]

[11] C. L. Lawson, H. Berman, L. Chen, B. Vallat, and C. Zirbel, “The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids,” Nucleic Acids Research, vol. 52, no. D1, pp. D245–D254, Jan. 2024. [Online]. Available: https://academic.oup.com/nar/article/52/D1/D245/7416380 [Cited on page 9.]

[12] M. Dumontier, “Aptamerbase.” [Online]. Available: https://github.com/micheldumontier/aptamerbase [Cited on page 9.]

[13] “Aptagen.” [Online]. Available: https://www.aptagen.com/apta-index/ [Cited on page 9.]

[14] D. Sun, M. Sun, J. Zhang, X. Lin, Y. Zhang, F. Lin, P. Zhang, C. Yang, and J. Song, “Computational tools for aptamer identification and optimization,” TrAC Trends in Analytical Chemistry, vol. 157, p. 116767, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165993622002503 [Cited on page 11.]

[15] B.-Q. Li, Y.-C. Zhang, G.-H. Huang, W.-R. Cui, N. Zhang, and Y.-D. Cai, “Prediction of Aptamer-Target Interacting Pairs with Pseudo-Amino Acid Composition,” PLoS ONE, vol. 9, no. 1, p. e86729, Jan. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3899287/ [Cited on pages 11, 26, and 47.]

[16] Q. Yang, C. Jia, and T. Li, “Prediction of aptamer–protein interacting pairs based on sparse autoencoder feature extraction and an ensemble classifier,” Mathematical Biosciences, vol. 311, pp. 103–108, May 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0025556418305923 [Cited on pages 12 and 26.]

[17] N. Emami and R. Ferdousi, “AptaNet as a deep learning approach for aptamer–protein interaction prediction,” Scientific Reports, vol. 11, no. 1, p. 6074, Mar. 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-85629-0 [Cited on pages 12, 26, and 47.]

[18] F. Morsch, I. L. Umasankar, L. S. Moreta, P. Latawa, D. B. Lange, J. Wengel, H. Konjen, and C. Code, “AptaBERT: Predicting aptamer binding interactions,” Nov. 2023. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.11. 24.568626v1 [Cited on page 12.]

[19] M. Zuker, “Mfold web server for nucleic acid folding and hybridization prediction,” Nucleic Acids Research, vol. 31, no. 13, pp. 3406–3415, Jul. 2003. [Cited on page 13.]

[20] Y. Zhao, Y. Huang, Z. Gong, Y. Wang, J. Man, and Y. Xiao, “Automated and fast building of three-dimensional RNA structures,” Scientific Reports, vol. 2, no. 1, p. 734, Oct. 2012. [Online]. Available: https://www.nature.com/articles/srep00734 [Cited on page 13.]

[21] Y. Zhang, Y. Xiong, and Y. Xiao, “3dDNA: A Computational Method of Building DNA 3D Structures,” Molecules, vol. 27, no. 18, p. 5936, Jan. 2022. [Online]. Available: https://www.mdpi.com/1420-3049/27/18/5936 [Cited on page 13.]

[22] C. Nithin, P. Ghosh, and J. M. Bujnicki, “Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes,” Genes, vol. 9, no. 9, p. 432, Aug. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6162694/ [Cited on page 14.]

[23] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, “HADDOCK: A ProteinProtein Docking Approach Based on Biochemical or Biophysical Information,” Journal of the American Chemical Society, vol. 125, no. 7, pp. 1731–1737, Feb. 2003. [Online]. Available: https://pubs.acs.org/doi/10.1021/ja026939x [Cited on pages 14 and 23.]

[24] H. Schweke, M.-H. Mucchielli, N. Chevrollier, S. Gosset, and A. Lopes, “SURFMAP: A Software for Mapping in Two Dimensions Protein Surface Features,” Journal of Chemical Information and Modeling, vol. 62, no. 7, pp. 1595–1601, Apr. 2022. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.1c01269 [Cited on pages 15 and 24.]

[25] S. A. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, and Z. Cao, “Efficacy of different protein descriptors in predicting protein functional families,” BMC Bioinformatics, vol. 8, no. 1, p. 300, Aug. 2007. [Online]. Available: https: //doi.org/10.1186/1471-2105-8-300 [Cited on page 15.]

[26] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, “From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment,” Cell, vol. 186, no. 8, pp. 1772–1791, Apr. 2023. [Cited on page 16.]

[27] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zˇ´ıdek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, no. 7873, pp. 583–589, Aug. 2021. [Cited on page 16.]

[28] T. K. Ho, “Random decision forests,” in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, Aug. 1995, pp. 278–282 vol.1. [Online]. Available: https://ieeexplore.ieee.org/document/598994 [Cited on page 17.]

[29] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” Psychological Review, vol. 65, no. 6, pp. 386–408, Nov. 1958. [Cited on page 17.]

[30] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in Proceedings of the fifth annual workshop on Computational learning theory, ser. COLT ’92. New York, NY, USA: Association for Computing Machinery, Jul. 1992, pp. 144–152. [Online]. Available: https://doi.org/10.1145/130385.130401 [Cited on page 17.]

[31] M. Silva, D. Pratas, and A. J. Pinho, “Efficient DNA sequence compression with neural networks,” GigaScience, vol. 9, no. 11, p. giaa119, Nov. 2020. [Online]. Available: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaa119/5974977 [Cited on page 18.]

[32] “[1603.02754] XGBoost: A Scalable Tree Boosting System,” Aug. 2016. [Online]. Available: https://web.archive.org/web/20160817055008/http://arxiv.org/abs/1603.02754 [Cited on page 18.]

[33] J. H. Friedman, “Stochastic gradient boosting,” Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, Feb. 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947301000652 [Cited on page 18.]

[34] S. Kumar, M. Nehra, J. Mehta, N. Dilbaghi, G. Marrazza, and A. Kaushik, “Point-of-Care Strategies for Detection of Waterborne Pathogens,” Sensors (Basel, Switzerland), vol. 19, no. 20, p. 4476, Oct. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6833035/ [Cited on page 19.]

[35] C. Manzanas, E. Morrison, Y. S. Kim, M. Alipanah, G. Adedokun, S. Jin, T. Z. Osborne, and Z. H. Fan, “Molecular testing devices for on-site detection of E. coli in water samples,” Scientific Reports, vol. 13, no. 1, p. 4245, Mar. 2023. [Cited on page 19.]

[36] R. V. Honorato, J. Roel-Touris, and A. M. J. J. Bonvin, “MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing,” Frontiers in Molecular Biosciences, vol. 6,Oct.2019.[Online].Available:
https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2019.00102/full [Cited on page 23.]

[37] J. Dong, Z.-J. Yao, L. Zhang, F. Luo, Q. Lin, A.-P. Lu, A. F. Chen, and D.-S. Cao, “PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions,” Journal of Cheminformatics, vol. 10, no. 1, p. 16, Mar. 2018. [Online]. Available: https://doi.org/10.1186/s13321-018-0270-2 [Cited on page 24.]

[38] A. C. Pereira, A. F. Pina, D. Sousa, D. Ferreira, C. Santos-Pereira, J. L. Rodrigues, L. D. R. Melo, G. Sales, S. F. Sousa, and L. R. Rodrigues, “Identification of novel aptamers targeting cathepsin B-overexpressing prostate cancer cells,” Molecular Systems Design & Engineering, vol. 7, no. 6, pp. 637–650, 2022. [Online]. Available: https://xlink.rsc.org/?DOI=D2ME00022A [Cited on page 24.]

[39] M. Silva, D. Pratas, and A. J. Pinho, “AC2: An Efficient Protein Sequence Compression Tool Using Artificial Neural Networks and Cache-Hash Models,” Entropy, vol. 23, no. 5, p. 530, May 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/5/530 [Cited on page 24.]

[40] P. Dua, S. Ren, S. W. Lee, J.-K. Kim, H.-S. Shin, O.-C. Jeong, S. Kim, and D.-K. Lee, “Cell-SELEX Based Identification of an RNA Aptamer for Escherichia coli and Its Use in Various Detection Formats,” Molecules and Cells, vol. 39, no. 11, pp. 807–813, Nov. 2016. [Cited on pages 29, 48, and 53.]