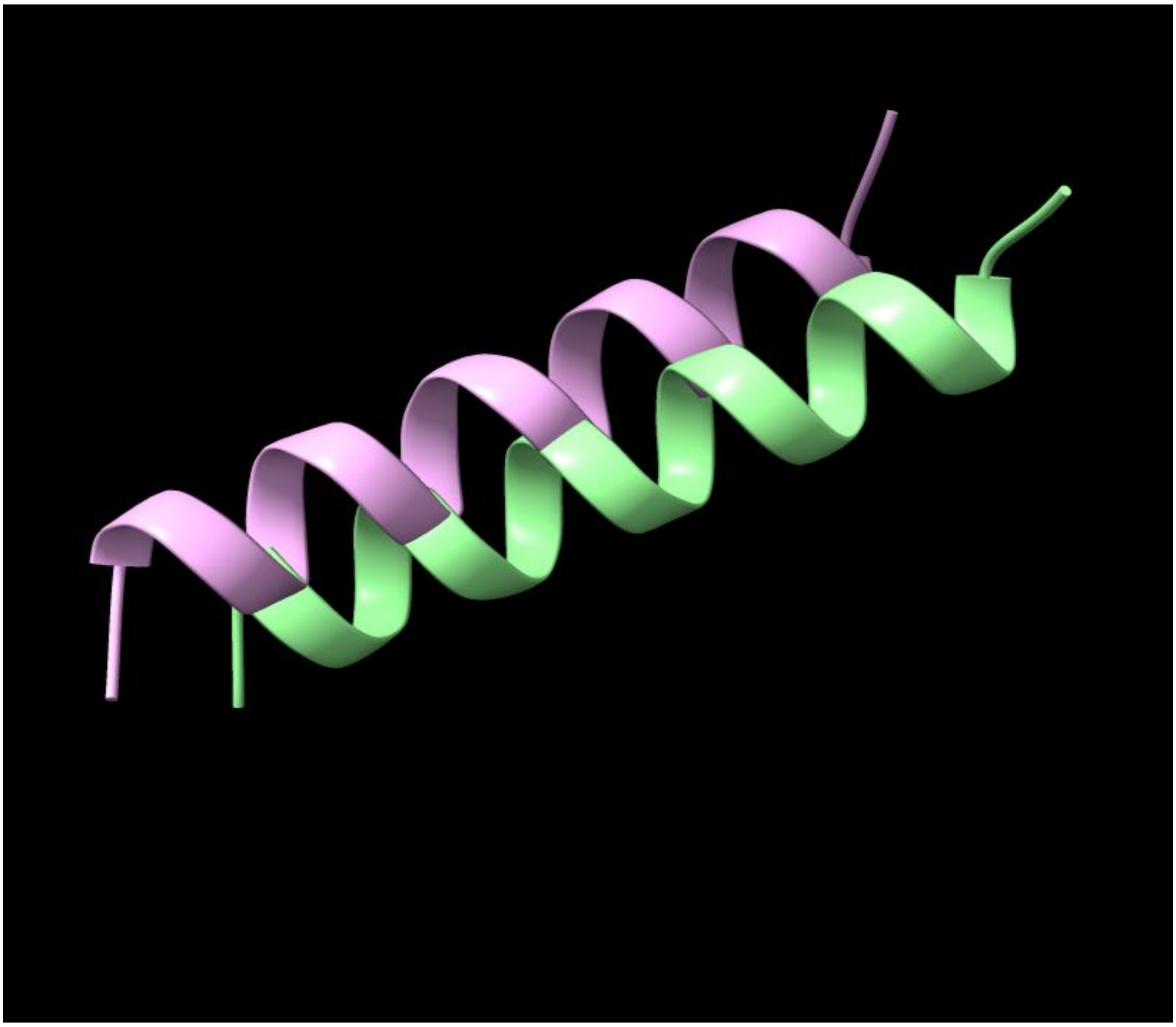
Unraveling the functions and implications of non-canonical microproteins in Neuroblastoma



Author: Amalia Nabuurs
Daily supervisors: Damon Hofman & Jip van Dinter
Examiner: Dr. Sebastiaan van Heesch
2nd examiner: Dr. Patrick van Kemmeren

Table of Contents

Layman’s summary

Neuroblastoma is a childhood cancer that accounts for 12-15% of childhood cancer deaths. The mechanism of disease and the development of neuroblastoma is not fully understood. This lack of knowledge and high mortality is a problem, and we would like to understand neuroblastoma better. Recent research suggests that microproteins (super small proteins) can have roles in cells. But we need to find microproteins with a function first out of thousands of microproteins that are found. We do this by combining data. DNA, the genomic information stored in all cells of our body, is the basis for RNA and proteins. DNA gets transcribed into RNA. You can see the DNA as a book with all blueprints for different building blocks within cells, and RNA as a copy of one of the blueprints. The RNA copy is used to build the protein. In all steps involved in this process, errors can occur, which sometimes lead to disease. To look for proteins that are specific to a tumor, you can examine the DNA (book), RNA (blueprints), and proteins and compare them to data from other tissues. While looking at DNA and RNA is getting easier and cheaper, studying proteins in cells remains a challenge. To discover new proteins built in a tumor, our lab has established a two-step protocol. The first part involves collecting all RNA in a specific tumor type, thereby collecting all the blueprints present in the tumor. The second part makes use of a new technique. In every cell, you have ribosomes, which are the machines that produce proteins based on the RNA blueprints. With this technique, only the RNA that is within the ribosome (machine) is examined. These are just small sections of the RNA that are protected within the ribosome, which is essentially parts of the blueprint. This is where the first step comes in. These small pieces are cross-referenced with the RNA to identify their corresponding RNA blueprints. With this technique, we can determine which proteins are made in the cell and discover new proteins. For a specific tumor type, thousands of these new proteins can be found. Out of these thousands of proteins, we must select candidates for further research. Examining all of them would be time and cost consuming. Interesting candidates have a function and need to be presented on the outside of the cell so the immune system can recognize and attack them. To automate this process, we have built a pipeline, a pipeline is a chain of several computational tools. These computational tools are built to predict properties of proteins. These properties are their structure, location in the cell, conservation and characteristics that can be calculated. We looked at proteins that were specific to neuroblastoma and enriched in neuroblastoma, and we found some promising proteins for further research based on the predictions of the pipeline.

Abstract

Microproteins, though often overlooked, have the potential to revolutionize cancer treatments, including immunotherapy and cellular therapies. With the advancements in genomics and proteomics thousands of potential microproteins are discovered. These microproteins could have crucial function in the tumor and possibly lead to several different new applications in cancer treatment. To identify translated microproteins our lab established a two-step protocol which combines making a custom transcriptome based on RNA-sequencing data of a tumor type, and mapping ribosome profiling data, which is based on the sequence that is protected in the ribosome, of the same tumor type against it. Two important implications of these microproteins are as target in immunotherapy or having a function in the tumor. Prior to targeting microproteins we need to do extensive experiments to validate their expression and functions in the tumor. To prioritize targets for further investigation, we have designed a pipeline that predicts several properties based on a list of microprotein sequences. Our pipeline gives information about the localization, structure, several characteristics, and MHC-I binding affinity of microproteins. We use the pipeline to specifically look at microproteins found in neuroblastoma. Neuroblastoma is a pediatric tumor that account for 12-15% of childhood cancer deaths. It originates somewhere in the sympathetic chain but the whole process is still unknown. This makes it a tumor type that would benefit a lot when tumor specific microproteins with a function are found. The pipeline gives us several targets which are enriched in neuroblastoma and have interesting properties. The pipeline also predicts several microproteins with potential strong binders with the MHC-I complex suggesting that there is potential to use them for immunotherapy. We come up with 3 candidates that are good candidates for further research.

Introduction

Microproteins are adding a new layer of complexity to our understanding of the genetic code. For years, the field has primarily focused on proteins, with efforts to map the entire human proteins, that were bigger than 100 amino acids. In most research an arbitrary cut-off of a hundred amino acids is used. This cut off limited the number of false positives proteins found and these small proteins were suspected to be noise or non-functional (Dinger et al. 2008). Recent studies indicate the contrary and implicate that microproteins may indeed have functions (Hassel, Brito-Estrada, and Makarewich 2023; Merino-Valverde, Greco, and Abad 2020). State-of-the-art genomics and proteomics also have led to the discovery of thousands of potential microproteins (Heesch et al. 2019)(Chothani et al. 2022). Microproteins are just one facet of a broader category of genetic elements known as Open Reading Frames (ORFs). ORFs are sequences within DNA that have the potential to be translated into proteins. There are several different types of ORFs depending on their location next to, or over canonical proteins. An ORF can be from annotated coding sequence (CDS) (fig. 1, left). But ORFs that are outside of the CDS or extensions of known ORFs are also possible. ORFs can be completely unannotated and novel, upstream ORFs (uORFs), downstream ORFs (dORFs), internal ORFs (intORFs), or overlapping uORFs (uoORFs) or overlapping dORFs (doORFs) (fig. 1, right) (Wright et al. 2022). These ORFs can vary widely in length, with some encoding conventional, well-established proteins and others, like microproteins, representing smaller and less-studied proteins. There is evidence that uncharacterized ORFs are functional (Prensner et al. 2021) and that investigating of unannotated ORFs in cancer and other disease states probably will yield new insights.

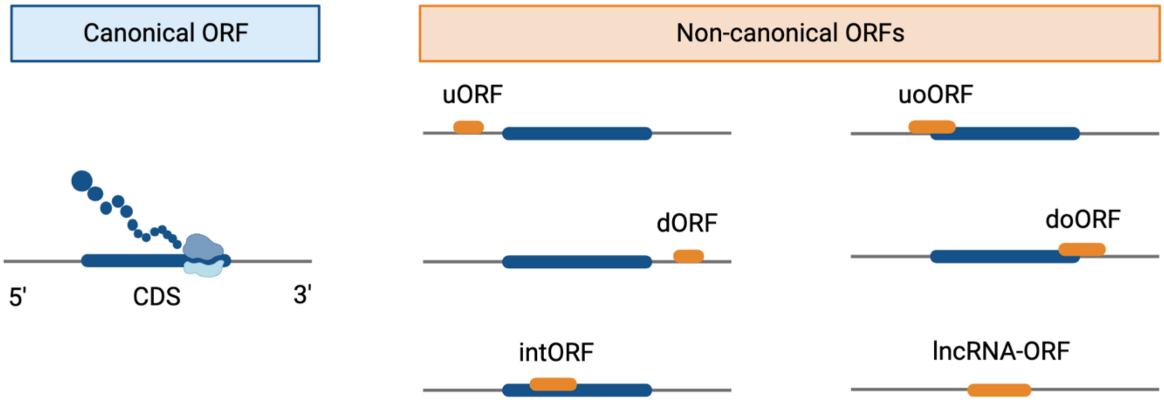


Figure 1: Several types of translated ORFs. In blue canonical ORFs are depicted. In orange non-canonical proteins are depicted. uORF = upstream open reading frame. dORF = downstream open reading frame. uoORF = upstream overlapping open reading frame. doORF = downstream overlapping open reading frame. intORF = internal open reading frame.

Proteins from these ORFs can be functional, regulate translation, be a source of novel proteins or be a source of novel antigenic peptides (fig. 2). A microprotein can have a function in complex stabilization, protein regulation, signaling or as an autonomous protein (fig. 2) (Schlesinger and Elsässer 2022). Microproteins can regulate translation, it can repress or upregulate a canonical ORF (fig. 2) (Schlesinger and Elsässer 2022). Microproteins can evolve over time and acquire a function especially in disease states where previously silent sections of DNA become transcriptionally active. Microproteins can be completely novel when they are from previously silenced DNA (fig. 2). Microproteins can be a source of antigenic peptides when the microproteins gets degraded and assembled on the MHC-I complex. This works by degrading a protein into peptides, these peptides with a length between 8 and 12 amino acids bind on the MHC-I (Blees et al. 2017). T-cells recognize the MHC-I and destroys cells when the assembled peptides are foreign. These possible functions and the presentation of antigenic peptides on the MHC-I complex make microproteins an interesting part of the proteome to study.

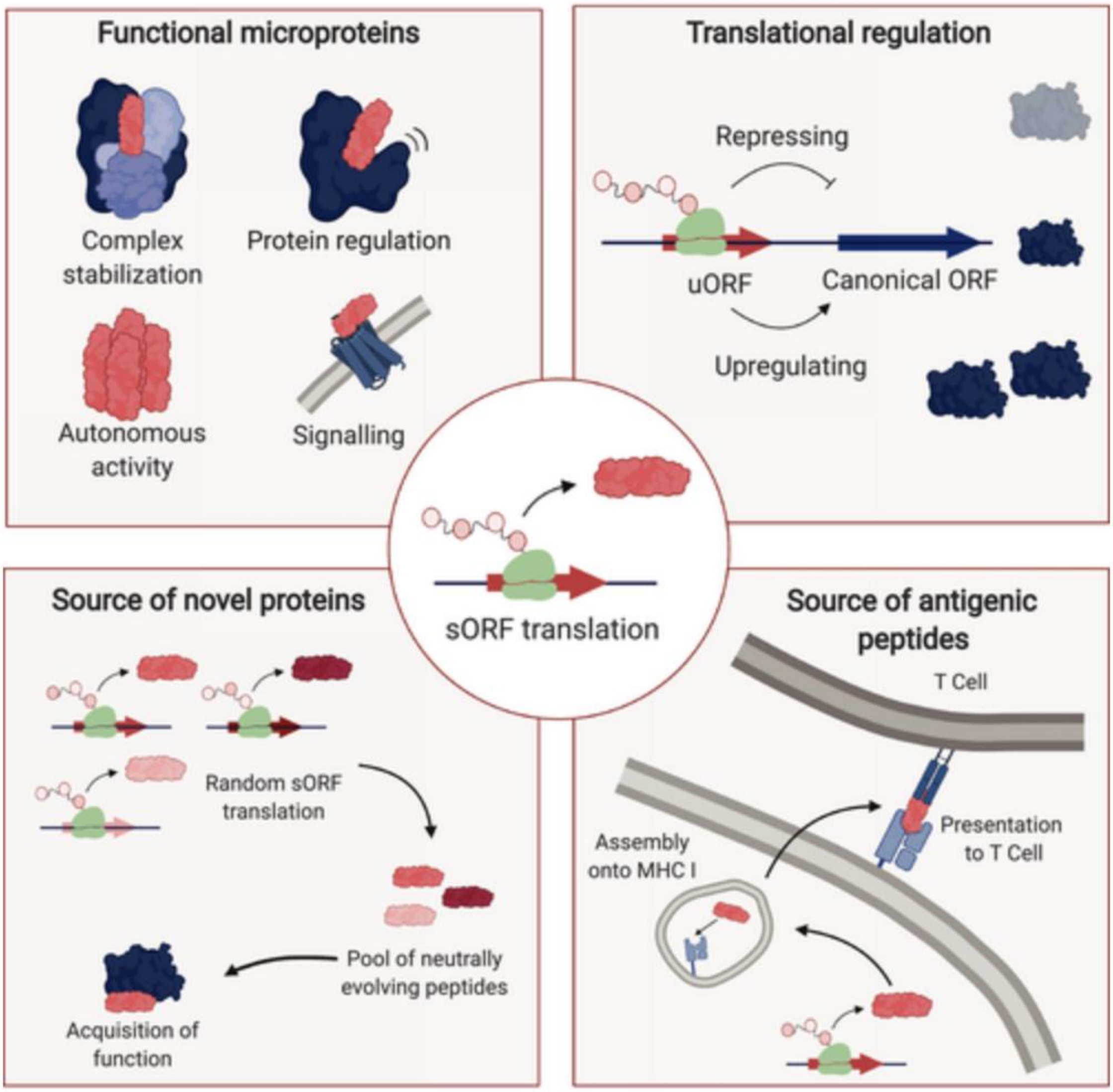


Figure 2: Small ORF translation and the potential function of the small ORF derived microproteins (Schlesinger and Elsässer 2022).

When novel peptides are presented by the MHC-I complex they could be potentially used for immunotherapy. These interventions include immune checkpoint inhibition, antibody-mediated therapy, and adoptive T cell therapy (Zhang and Zhang 2020). With immunotherapy, the body’s immune system is harnessed to recognize and destroy cancer cells, often through the identification of cancer-specific peptides which act as unique markers on tumor cells, enabling targeted immune responses. Immunotherapy options are already available for several cancer types, and this field is rapidly advancing, offering new ways of treatment for some types of cancer (Abbott and Ustoyev 2019). Most immunotherapies primarily target mutated canonical proteins in adult cancer types, where the proteins have gained the mutations over time. This approach has limitations in the case of pediatric cancer due to their low mutational burden (Filbin and Monje 2019), which makes novel microproteins a more suitable and promising candidate for immunotherapy in pediatric cancer.

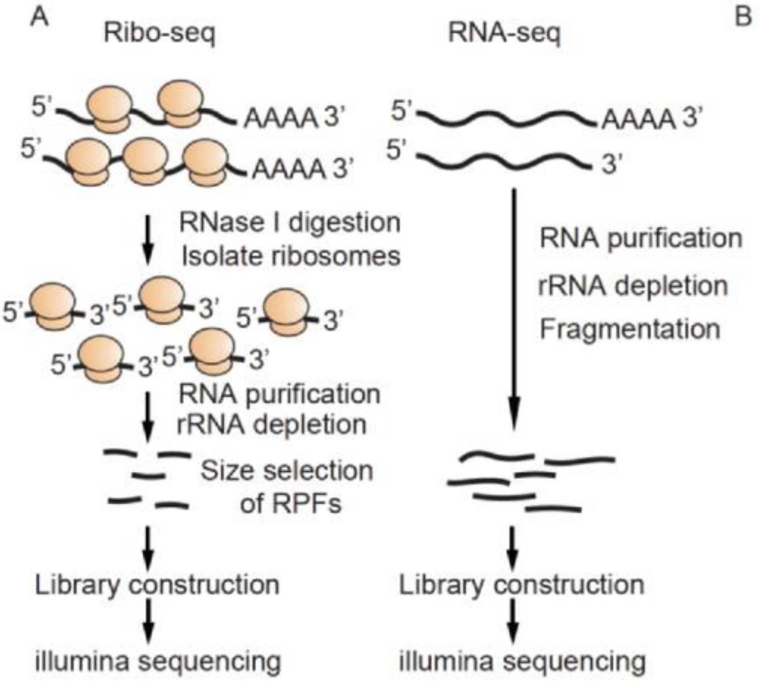


Figure 3: Overview of Ribo-seq and RNA-seq. A) on the left the pathway of ribo-seq, it starts at the top with digestion, then purification, size selection, library construction and at the end illumine sequencing. B) on the right RNA-seq pathway, it starts at the top with RNA purification, depletion, and fragmentation, after this library construction and illumine sequencing is performed (Xu et al. 2020).

Both functional and novel microproteins could be of interest for further research, but they need to be identified first out of thousands of possible translated microproteins. To identify microproteins we have established a two-step protocol. This protocol is performed to identify ORFs that are translated in a tissue-specific manner. The protocol makes use of ribosome profiling (Ribo-seq) and RNA sequencing to find evidence for actively translated ORFs (fig. 3). RNA sequencing (RNA-seq)(fig 3B) is a high-throughput molecular biology method that enables the analysis of the transcriptome, the complete set of RNA molecules present in a cell or tissue at a given moment (Withanage, Liang, and Zeng 2022). This technology has revolutionized the ability to identify novel genes, quantify their expression levels, and detect various RNA isoforms and alternative splicing events. RNA sequencing allows us to explore the transcriptomic landscape of a tissue. Combining the RNA-seq data of several samples of a tissue, a tissue specific transcriptome is built. This tissue specific transcriptome can be used to map ribosome profiling data against. Ribosome profiling shows the process of translation, the translation of RNA into functional proteins. Ribosome profiling (fig 3A), also known as Ribo-seq, sequences only the mRNA encapsulated in the ribosome (Ingolia et al. 2009). It involves capturing the positions of ribosomes along messenger RNAs (mRNAs) at a genome-wide scale. This technique allows to discern which mRNAs are actively being translated and provides a quantitative measure of ribosome occupancy on individual transcripts. By mapping the ribosome profiling data to the assembled tissue-specific transcriptome, we know which specific mRNAs are actively being translated into proteins ultimately giving us a list of translated ORFs.

This list of ORFs contains possible new microproteins, but this list contains all translated ORFs which are too many to investigate individually. And there is no evidence that the microproteins on the list are in fact stable or have a function. The development of new methodologies and advances in omics technologies have enabled large-scale discovery of previously unannotated ORFs. One of the next challenges is to further annotate, characterize, and validate the function of their candidate microprotein products. Using state of the art bioinformatics tools we are going to try and solve these challenges and predict several properties of novel microproteins. These predictions will give a better understanding about the microprotein and their possibility to have a crucial role in the tumor or to be used as target for immunotherapy. The pipeline can thus be used to prioritize microproteins for functional experiments.

The pipeline is built to address several areas of interest regarding microproteins:

- Cellular Localization: Determining where microproteins localize within cells provides insights into their functional roles. Microproteins may play distinct roles depending on their subcellular localization, influencing cellular processes and interactions.
- MHC-I binding affinity: Investigating the binding affinity of microproteins to Major Histocompatibility Complex Class I (MHC-I) molecules provides a critical perspective on their potential involvement in antigen presentation and probability to be used for immunotherapeutic interventions.
- Structural Analysis: Their three-dimensional structures, offer a deeper insight into their functional potential and interactions with other biomolecules.
- Function: Uncovering the functional roles of microproteins is important. These roles may include regulatory functions in cancer progression, immunomodulation, or roles in other cellular processes.
- Conservation: Conservation in microproteins would mean selection, selection means more likely to have a function. Microproteins without conservation are still of interest because microprotein are evolutionary young.
- Characteristics: defining several characteristics of microproteins would be interesting when comparing them to canonical microproteins to find similarities and differences between the sets.
- Short Linear Motif (SLiM) search: Intrinsically disordered proteins can have functions and interactions based on short linear motifs. These motifs can give us more information about the disordered microproteins.

By analyzing these aspects, we try to find out the roles that microproteins play and their potential to be used in immunotherapy, ultimately advancing our research on tumor specific microproteins.

As a first use case of the pipeline, we use Neuroblastoma ORF data with canonical and non-canonical microproteins. Annually, in the Netherlands, around 25 children get diagnosed with neuroblastoma (Neuroblastoom n.d.). Neuroblastoma which is a pediatric tumor that affects the sympathetic nervous system arises from neural crest progenitor cells. The prognosis of neuroblastoma heavily relies on its disease stage. The chances of recovery vary significantly, ranging from 70-90% to 25-50% based on age and risk status. Where low risk has a favorable prognosis with a 5-year overall >90% survival (Wienke et al. 2021). High risk neuroblastoma has a below 50% 5-year overall survival (Matthay et al. 2016). Early age of onset (3-5 years), high frequency of metastatic disease at diagnosis and tendency of spontaneous regression of tumors in infancy (Matthay et al. 2016) and the high risk variant of neuroblastoma sparked the interest in alternative therapeutic approaches. Neuroblastoma also displays low immunogenicity due to its low mutational load and lack of MHC-I expression (Wienke et al. 2021). The relatively low survival rates for neuroblastoma combined with the low immunogenicity indicate that patients would benefit from an alternative way of treatment.

Material & Methods

Software and algorithms

Table 1: Overview of used software, versions with their source and identifier.

Name	Version	Source	Identifier
chimeraX	1.5	(Meng et al. n.d.)	https://doi.org/10.1002/pro.4792
Rstudio	4.3.0	R Core Team, 2021	https://www.R-project.org/
Python	3.7	(Van Rossum, et al. 2009)	https://www.python.org/
OmegaFold	Model 2	(Wu et al. 2022)	https://doi.org/10.1101/2022.07.21.500999
DeepTMHMM	1.0.24	(Hallgren et al. 2022)	https://doi.org/10.1101/2022.04.08.487609
SignalP	6.0	(Teufel et al. 2022)	https://doi.org/10.1038%2Fs41587-021-01156-3
Peptides	2.4.5	(Osorio et al. 2023)	https://github.com/dosorio/Peptides/
IUPred	3.0	(Erdős, Pajkos, and Dosztányi 2021)	https://doi.org/10.1093%2Fnar%2Fgkab408
NetMHCpan	4.1	(Reynisson et al. 2020)	https://doi.org/10.1093/nar/gkaa379
Ggplot2	3.3.5	(Wickham, Chang, et al. 2023)	https://ggplot2.tidyverse.org
tidyr	1.2.0	(Wickham, Vaughan, et al. 2023)	https://tidyr.tidyverse.org
pandas	1.1.5	The pandas development team	https://doi.org/10.5281/zenodo.8364959
dplyr	1.0.8	(Wickham, François, et al. 2023)	https://dplyr.tidyverse.org
BioStrings	2.70.1	(Pagès, et al. 2023)	https://doi.org/doi:10.18129/B9.bioc.Biostrings
stringr	1.4.0	(Wickham and RStudio 2022)	https://stringr.tidyverse.org
magrittr	2.0.3	(magrittr) et al. 2022)	https://magrittr.tidyverse.org
numpy	1.24	(Harris et al. 2020)	DOI: 10.1038/s41586-020-2649-2

TrimGalore	0.6.6	(Kreuger F, 2021)	https://zenodo.org/badge/latestdoi/62039322
Cutadapt	4.6	(Martin 2011)	https://doi.org/10.14806/ej.17.1.200
FastQC	1.5.0	(Andrews S. 2010)	https://doi.org/10.5281/zenodo.6984534
StringTie	2.1.5	(Pertea et al. 2015)	https://doi.org/10.1038%2Fnbt.3122
STAR	2.7.8a	(Dobin et al. 2013)	https://doi.org/10.1093/bioinformatics/bts635
Bowtie	2	(Langmead and Salzberg 2012)	https://doi.org/10.1038%2Fnmeth.1923
VennDiagram	1.7.1	(Chen 2022)	https://CRAN.R-project.org/package=VennDiagram

Data and scripts availability

Scripts

All scripts used in this study can be found on the following git repository: https://github.com/AmaliaNabuurs/protein_prediction_project.git

Containers

All tools used are containerized to ensure easy reproducibility. Dockers created for this project can be found at dockerhub: https://hub.docker.com/u/anabuurs. Their corresponding dockerfiles are all in the following git repository: https://github.com/AmaliaNabuurs/dockerfiles.git. The only exception is the docker for IUPred3.0, this one is set on private due to licensing. The code for IUPred3.0 can be requested on https://iupred3.elte.hu/download_new.

Existing dockers:

- netMHCpan - https://hub.docker.com/r/guobioinfolab/netmhcpan
- SignalP 6.0 - https://hub.docker.com/r/streptomyces/signalp

Visualization of PDB files from OmegaFold was done with chimeraX 1.5. Molecular graphics and analyses performed with UCSF chimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Data

Total RNA sequencing data on neuroblastoma samples (n=233) were collected from the Princess Maxima Centre for Pediatric Oncology. Total Ribosome profiling data on neuroblastoma samples (n=15) were collected from the Princess Maxima Centre for Pediatric Oncology.

Data pre-processing

Neuroblastoma transcriptome creation

Creating a neuroblastoma specific transcriptome was done based on 233 RNA sequencing samples. Quality control and trimming was done with Trim Galore 0.6.6, a wrapper around Cutadapt and FastQC. STAR 2.7.8a was used to align the reads to the genome, guided by transcriptome Ensembl v102. Stringtie 2.1.5 was used to assemble the transcriptome. All these steps were performed as in our lab's protocol.

Translatome creation

Creating a neuroblastoma translatome was done based on 15 neuroblastoma ribosome profiling samples. Quality control and trimming was performed by Trim Galore. Contaminant RNA and DNA were removed using Bowtie2. The unmapped ribosome protected fragments were aligned using STAR to the custom neuroblastoma transcriptome. All identified ORFs were merged and ORFs which appeared in at least 2 samples were kept. All steps were performed as in our lab's protocols.

Identification of non-canonical ORFs

ORFquant detects and quantifies ORF translation on complex transcriptomes using Ribo-seq data. It quantifies translation at the single ORF level taking into account the presence of multiple transcripts expressed by each gene. It outputs info into a GTF file.The GTF from ORFquant was used to extract information about the ORFs and filter them for neuroblastoma enriched and specific ORFs. This was done by filtering on ORF type keeping only dORFs, novel ORFs, uORFs, and lncORFs. And filtering on class type removing all ORFs with class k, as most of the novel ORFs derived from k-class transcripts very closely resembled canonical CDS regions. After all the ORFs in neuroblastoma were detected, they were filtered on several criteria. To get non-canonical ORFs filtering was based on known transcript and proteins from the UniPROTdb & GTEx.

Resulting neuroblastoma datasets

Three datasets based on 15 neuroblastoma ribo-seq samples:

1. Neuroblastoma non-canonical microproteins – 2464 microproteins.
2. Neuroblastoma canonical microproteins – 3120 microproteins.
3. Neuroblastoma enriched non-canonical microproteins – 25 microproteins.

All data was filtered for microproteins with a length between 19 and 150 amino acids.

Creation of pipeline

Selection of tools

The pipeline has been developed to predict various aspects related to microproteins. For each specific category of interest, extensive literature research was conducted to identify the most fitting prediction tools. This section will provide an in-depth discussion of the selected tools for each category, providing the rationale behind their selection. Key considerations in evaluating these tools include their methodology, performance on microproteins, input and output file requirements, advantages, disadvantages, and strategies for optimizing their utilization on our cluster infrastructure.

Selecting 3D structure predictor

The three-dimensional structures of microproteins, offer a deeper insight into their functional potential and potential interactions with other biomolecules (Lu, Fornili, and Fraternali 2013). To find a structure predictor that suits our research question we looked at the most widely used and promising structure predictors that are currently available and compared them. It is worth noting that microproteins are often characterized by a lack of significant conservation across species, necessitating their treatment as *de novo* or orphan proteins due to the absence of conservation data (Sandmann et al. 2023). Most structure predictors are within two groups; 1. proteins langue models or 2. multiple sequence alignment (MSA) based models which also use trained neural networks. The MSA based models, Alphafold2, OpenFold, and MODELLER (table 2) perform good on canonical proteins but lack the power to predict *de novo* proteins (Ahdritz et al. 2022; Skolnick et al. 2021; Webb and Sali 2016). Proteins language based models, ESMFold, OmegaFold, and RGN2 (table 2) perform better on orphan or *de novo* proteins, making them particularly well-suited for our research needs (Chowdhury et al. 2022; Verkuil et al. 2022; Wu et al. 2022). Comparison of RGN2 and OmegaFold based on literature showed that OmegaFold performed better on *de novo* proteins than RGN2 (Aubel, Eicholt, and Bornberg-Bauer 2023). Consequently, we have chosen OmegaFold as the 3D structure predictor in our pipeline (Wu et al. 2022).

Table 2: Overview of several tools that can be used for visualization of 3D structure of microproteins.

Name	*Method*	*Input / output*	*Disadvantages*	*Advantages*	*Literature*
Alphafold2	MSA + neural network	Input: fasta Output: pdb	MSA based	Best protein predictor in the field star of CASP14	(Skolnick et al. 2021)
Openfold	MSA + neural network	Input: fasta output: pdb	Not trained less data than alphafold2	trained like alphafold2 but with open code and data access	(Ahdritz et al. 2022)
Omegafold	language based model + deep learning	Input: fasta Output: pdb	-	better when there is no evolutional info present	(Wu et al. 2022)

				- best pLM in comparison paper	
RGN2	language based model + deep learning	Input: fasta output: pdb	slightly worse than OmegaFold on tested proteins	Better for orphan and de novo proteins	(Chowdhury et al. 2022)
ESM 2 / ESMFold	language based model + deep learning	Input: fasta Output: pdb	Based on their own study performs worse than AlphaFold2 on de novo proteins	Better for de novo proteins	(Verkuil et al. 2022)
MODELLER	comparative modelling of protein three-dimensional structures spatial restraints	Input: alignment + atomic coordinates + script Output: several options	MSA based	- whole code and method are online - can do de novo proteins - is used in de novo microprotein studies	(Webb and Sali 2016)

OmegaFold is a cutting-edge protein language model that leverages deep learning techniques specifically designed for *de novo* protein structure prediction. This state-of-the-art model employs a neural network architecture to make accurate predictions about the three-dimensional structure of proteins, providing valuable insights into their spatial arrangement and conformation. For each protein analyzed, OmegaFold generates a PDB (Protein Data Bank) file containing the precise coordinates of all its amino acids. These PDB files can be utilized for visualizing the protein structures using software applications such as pyMOL and ChimeraX (Pettersen et al. 2021). OmegaFold provides a crucial assessment metric in the form of a pLDDT (predicted Local Distance Difference Test) score for each residue within the predicted structure. This score serves as an indicator of the confidence level associated with each prediction. A breakdown of how to interpret these pLDDT scores is here:

1. **pLDDT between 70 and 90:** These regions are expected to be modelled with high accuracy, reflecting a generally reliable backbone prediction. Researchers can have confidence in the structural information within this range.
2. **pLDDT between 50 and 70:** In regions where pLDDT falls within this range, caution is advised. These areas are considered low confidence, indicating potential uncertainties or deviations in the structure. Further validation or scrutiny may be necessary in these regions.
3. **pLDDT < 50:** Regions with pLDDT scores below 50 should be approached with extreme caution. The 3D coordinates in these areas often exhibit a ribbon-like appearance, and their structural interpretation should be avoided due to the significant level of uncertainty associated with these predictions.

In summary, OmegaFolds utilization of the pLDDT score not only aids in assessing the reliability of its predictions but also assists in distinguishing between well-modelled regions and those requiring more scrutiny or validation. This information is instrumental in guiding the interpretation and application of the predicted protein structures in downstream analyses and experiments.

Selecting Disorder prediction tool

The absence of well-defined protein structures often signifies disorder, which refers to the lack of a stable tertiary structure. For microproteins, it is well-established that these typically possess a small functional domain while exhibiting a significant degree of disorder throughout their structure (Wilson et al. 2017). Disordered regions are evolutionary less conserved compared to ordered regions (Erdős, Pajkos, and Dosztányi 2021). Important for disorder prediction is the DisProt database which includes experimentally verified disordered segments (Quaglia et al. 2022).

In our evaluation, we compared three disorder predictors: IUPred 3.0, flDPnn, and SPOT-Disorder2. Based on the results of reviews performed by others and insights gathered from the Critical Assessment of protein Intrinsic Disorder (CAID) challenge suggest that flDPnn is a top-performing predictor for disorder (Aubel, Eicholt, and Bornberg-Bauer 2023; Conte et al. n.d.; Necci et al. 2021; Quaglia et al. 2022). However, flDPnn encountered technical issues and was unable to function effectively on our cluster infrastructure. Upon reviewing and reevaluating the results from the CAID challenge, we determined that IUPred3 offers a nearly equivalent performance to flDPnn in predicting disorder. Notably, IUPred3 demonstrated compatibility with our cluster environment, making it a feasible choice for integration into our pipeline. Moreover, within the protein research community, IUPred 3.0 has gained prominence and is frequently employed (Sandmann et al. 2023; Schmitz, Ullrich, and Bornberg-Bauer 2018; Wilson et al. 2017; Xie et al. n.d.). In light of these considerations, we have adopted IUPred 3.0 as the primary tool for disorder prediction in our pipeline, ensuring the accurate assessment of disorder within microproteins and enhancing the robustness of our analyses.

Table 3: Overview of disorder predictors.

	Method	*Input / output*	*Dis advantages*	*Advantages*	*Literature*
IUPred3	Energy estimation method	input amino acid sequence output text, graphical and json	Download possible via asking/email for academic	Most widely used disorder predictor	(Erdős, Pajkos, and Dosztányi 2021)
flDPnn	neural network + random forest	input fasta output txt & visualizations	Does not work on cluster	docker available best according to CAID	(Hu et al. 2021)
SPOT-Disorder2	long short term memory (LSTM) and excitation residual inception	input fasta file output text file	Performs worse than flDPnn and IUPred	downloadable package	(Hanson et al. 2019)

IUPred 3.0 is based on a unique energy estimation approach that provides fast and robust prediction of disordered tendency. The output from IUPred 3.0 includes a file for each protein, containing disorder scores assigned to all individual residues. When the disorder score surpasses the threshold of 0.5 for a given residue, it is flagged to be in a disordered state. An average is calculated for all input proteins. In addition to the average disorder score,

comma-separated values for the disorder scores per residue are included in the final table. This granular level of information is valuable for pinpointing specific regions of disorder within each protein, as variations can be substantial from one region to another. Furthermore, it is important to mention that IUPred 3.0 imposes a minimum protein length requirement of 19 amino acids for accurate prediction. This ensures that the tool operates effectively and reliably, as shorter sequences may not provide sufficient information for disorder prediction.

Short linear motif (SLiM) search

In microproteins characterized by a substantial degree of disorder, predicting their function can be challenging due to their lack of a well-defined structure. However, it is possible to infer potential functions or anticipate protein-protein interactions by examining Short Linear Motifs (SLiMs) (Van Roey et al. 2014). SLiMs are typically short peptide sequences, ranging from 3 to 15 amino acids in length, with 2 to 5 defined positions. They are known to occur by chance and can be difficult to identify (Edwards and Palopoli 2015).

To facilitate the discovery of SLiMs within microproteins, there are several specialized resources and tools available. The Eukaryotic Linear Motif (ELM) resource is the most comprehensive repository of experimentally validated SLiMs (Van Roey et al. 2014). HH-Motif employs Hidden Markov Models to identify SLiMs within protein sequences. It is available both as a webserver and a standalone version. However, one limitation of HH-motif is that it can only detect SLiMs in proteins larger than 50 amino acids, which is not suitable for the microproteins due to their small size. QSLiMFinder is a short linear motif predictor using specific query protein data. It uses a statistical model, SLIMChance, to calculate the probability of a SLiM. The MEME suite, a Motif-based sequence analysis toolbox, can find Motifs you provide in protein sequences. SLiMPred computationally predicts SLiM regions in protein sequences. It uses machine learning methods to find motifs that are in the Eukaryotic Linear Motif database. All tools created for SLiM finding that we took into consideration are web-based tools. Because of this we decided to write a script that searches for the regex sequences that are in the Eukaryotic Linear Motif database ourselves. Tools that use machine learning perform better but we decided to use a regex which is robust and informative enough at this point.

Table 4: Overview of SLiM search algorithms that are compared.

Name	*Method*	*Disadvantages*	*Advantages*	*Literature*
ELM search	Regex search	Regex has no interdependencies between sides	Easy to use	(Kumar et al. 2022)
QSLiMFinder	Query protein data	Correct for evolutionary relationships	-	(Palopoli, Lythgow, and Edwards 2015)
HH-Motif	Hidden Markov model	Can only find SLiM in protein >50 amino acids	-	(Prytuliak et al. 2017)
MEME	Regex search	Needs list of SLiMs as input	-	(Bailey et al. 2015)
SLiMPred	Machine learning method	Only a webserver	-	(Mooney et al. 2012)

For the detection of SLiMs the ‘elm_classes.tsv’ file (15 July 2023) was downloaded from the ELM resource. This file contains the ELM classes they have a regex to define the possible amino acid sequence. All these regexes are sequentially searched. We then filtered the peptide sequences for matches to any of the motifs falling in regions with a disorder value ≥ 0.5. This left us with the ELM classes and sequences for the whole protein and for only the disordered parts in the protein.

Selecting localization tools

Determining the subcellular localization of microproteins is pivotal in unraveling their functional roles within cells. Microproteins can assume distinct functions based on their subcellular localization, exerting influence over cellular processes and interactions between proteins. Additionally, the identification of transmembrane segments and signal sequences increases our understanding of a proteins potential cellular localization.

To comprehensively investigate computationally the subcellular localization and structural attributes of microproteins, we explored various categories of tools:

1. Tools for Whole Protein Localization:

DeepLoc 2.0: Utilizes a neural network to predict the most likely subcellular localization of a protein, encompassing 10 areas within the cell. These are the Nucleus, Cytoplasm, Extracellular, Mitochondrion, Cell membrane, Endoplasmic reticulum, Chloroplast, Golgi apparatus, Lysosome/Vacuole and Peroxisome. It assigns a score to each possible location, and if the score surpasses a threshold, the protein is predicted to likely be localized there.

Quick2D: This web tool combines several predictive components, like the pipeline we aim to develop. It integrates multiple tools for predicting various sequence features, including secondary structure, intrinsically disordered regions, transmembrane regions, signal peptides, and coiled-coil regions. It provides a summary of information from different sources but has limitations regarding tool versions and the presentation of detailed information.

2. Tools Predicting Specific Features:

DeepTMHMM: Employs hidden Markov models (HMM) to predict transmembrane segments within proteins. It can predict both alpha helixes and beta sheets.

SignalP 6.0: Utilizes a neural network to detect the presence of signal peptides in protein sequences. In eukaryotic mode it predicts whether there is a signal present and where the signal is located in the sequence.

NetGPI 1.1: Utilizes a neural network to predict the presence of Glycosylphosphatidylinositol (GPI) anchors to the membrane in peptide sequences, which is particularly relevant if a protein is secreted.

TargetP 2.0: Predicts the presence of target signals directing proteins to the mitochondrion or plastids. Tools is primarily built for plant target peptides.

SamCC: Predicts the presence of coiled-coil regions within a protein sequence.

3. Tools Combining Predictions:

MembraneFold: Combines the predictions of DeepTMHMM and OmegaFold to deduce the structure of transmembrane segments in proteins.

Phobius: A comprehensive tool that predicts both transmembrane domains and signal peptides in protein sequences.

While Quick2D offers a good approach, it has limitations, such as its reliance on older version of TMHMM and the lack of detailed information presentation (DUAN et al. 2021; Gabler et al. 2020). For instance, it may not specify the origin of the signal detected by SignalP and may not provide the topology of membrane segments, which Phobius and TMHMM offer.

Additionally, Quick2D is limited in scalability as it can process only one microprotein at a time. MembraneFold and Phobius are tools that predict two properties of proteins at once. Since, these tools combine two tools, and leave information out, we decided to start with the tools that predict one feature of microproteins (Gutierrez et al. 2022; Käll, Krogh, and Sonnhammer 2004).

NetGPI, DeepTMHMM, SignalP, targetP and SamCC predict specific domains in proteins (Armenteros et al. 2019; Gíslason et al. 2021; Hallgren et al. 2022; SamCC-Turbo n.d.; Teufel et al. 2022). All tools are of interest to get to know something about microproteins. We started with SignalP and DeepTMHMM because these tools give two distinct predictions about proteins which we suspect to be present. TargetP, NetGPI and SamCC can be interesting to add later or revise on a later point. For the first version of the pipeline the focus is on DeepTMHMM and SignalP.

Table 5: tools for localization of (part of) the microprotein.

Name	*Method*	*input / output*	*advantages*	*disadvantages*	*literature*
Deep TMHMM	deep learning algorithm that uses a hidden markov model	input = fasta output = summary 3line and .md	Prediction of transmembrane parts	is also in the Quick2D toolkit in BioLib, hard to containerize	(Hallgren et al. 2022)
SignalP 6.0	Neural network	input = fasta output = text based	Predict signal peptides	-	(Teufel et al. 2022)
DeepLoc 2.0	Trained transformer language model	input = fasta output = text based	subcellular localization	Does not work on cluster	(Thumuluri et al. 2022)
Quick2D	Runs several tools to predict sequence features	input = protein sequence output = web interface	several tools at ones and puts them in a single graph with the outcomes	Not all information of the tools used is depicted. Some tools outdated	(DUAN et al. 2021; Gabler et al. 2020)
Phobius	A combined transmembrane topology and signal peptide predictor	input = protein sequence output = web interface	A combined transmembrane topology and signal peptide predictor	signal P and deeplock / TMHMM are better	(Käll, Krogh, and Sonnhammer 2004)
NETGPI 1.1	Deep learning approach	input = fasta output = text based	predicts whether protein has GPI anchors	Very specific	(Gíslason et al. 2021)
TargetP 2.0	Deep learning	Input = protein sequence Output = text based	Predicts target peptides to mitochondrion and plastids	Very specific, plant or non-plant	(Armenteros et al. 2019)
Membrane Fold	deepTMHMM + OmegaFold	input = protein sequence	two tools at once (deepTMHMM and	The tools independent are more interesting	(Gutierrez et al. 2022)

		output = text based	OmegaFold combined)		
SamCC	used to detect coiled coil domains (cc) (coiled coil = 2 or more alpha helixes)	Input = PDB coordinates Output = Text based	is for coiled coils specific	found in paper that looked at de novo microproteins	(SamCC-Turbo n.d.)

In the development of our pipeline, we have chosen three key tools to facilitate the analysis of microprotein localization: DeepTMHMM, SignalP, and DeepLoc. Each tool serves a distinct purpose in enhancing our understanding of microprotein characteristics. Together, they give the broadest detection of localization signals. These three tools have been selected to provide a comprehensive analysis of microproteins, encompassing aspects of structural features, localization, and their potential to be signaled to other cell compartments or secreted. Unfortunately, after multiple attempts we were not able to run DeepLoc locally, so it was decided to only use the tool only on a subset of microproteins that are deemed interesting based on predictions made by other tools.

Predicting MHC-I binding affinity

Investigating the binding affinity of peptides from the microproteins to Major Histocompatibility Complex Class I (MHC-I) molecules provides a critical perspective on their potential involvement in antigen presentation and as target for immunotherapeutic interventions. Based on previous research performed by the group which involved a comparison between several types of MHC-I binding affinity predictors netMHCpan was chosen.

NetMHCpan uses state of the art tailored machine learning strategies to integrate different training data types (Reynisson et al. 2020). This tool accepts input in the form of FASTA protein sequences and provides predictions for each possible peptide within the sequence, typically ranging from 8 to 12 amino acids in length. Specifically, it predicts the binding affinity of these peptides to a range of HLA (Human Leukocyte Antigen) molecules, the human equivalent of MHC. The twelve most common HLA types are included in the search: HLA-A01:01, HLA-A02:01, HLA-A03:01, HLA-A24:02, HLA-A26:01, HLA-B07:02, HLA-B08:01, HLA-B27:05, HLA-B39:01, HLA-B40:01, HLA-B58:01, and HLA-B15:01.

Microprotein conservation

Conservation in proteins means selection pressure which hints towards importance for cell survival. This means that conservation in microproteins would mean that the microproteins more probable to have a function. On the other hand, microproteins with homologs are less likely to be suitable as targets for immunotherapy. It is worth noting that microproteins are not expected to exhibit strong conservation across species, as they are known to be evolutionarily young (Sandmann et al. 2023). Conservation or lack of conservation in microproteins can be used to find out whether we would be interested in a microprotein because of function or because they are a candidate for immunotherapy. For conservation search three methods were investigated BLASTp, HMMER and HHpred.

BLASTp searches the NCBI BLAST database (Altschul et al. 1997). It is used to find homologs or orthologs of proteins. It performs a protein versus protein search and uses penalties when they do not align. Cut offs are used to determine whether a protein is evolutionary related or not. HMMER is a toolbox which can be used to search for distant evolutionary relationships. It uses a Hidden Markov Model which compares the protein to a protein database. HHpred is a protein function and protein structure prediction server that is based on HHsearch and HHblits, another program in the HH-suite package (Steinegger et al. 2019).

Table 6: Overview of conservation search tools.

Name	*Method*	*Input / output*	*Disadvantages*	*Advantages*	*Literature*
BLASTp	Protein vs protein search	In: fasta Out: summary text based	Hard for microproteins	Detects closest evolutionary relationships	(Altschul et al. 1997)
HMMER	Profile HMM to protein search	In: profile based on sequence Out: text based	Too far away evolutionary	Can detect more distant evolutionary relationships	(Potter et al. 2018)
HHpred	MSA + profile search	In: single sequence Out: text based	Too far away evolutionary	Can detect more distant evolutionary relationships	(Söding, Biegert, and Lupas 2005)

In our pipeline, BLASTp was initially configured to interact with the NCBI BLAST online database, which functioned adequately for smaller protein lists. However, when dealing with our final dataset, consisting of over 2000 microproteins, this approach resulted in excessive CPU usage on the NCBI BLAST online database. To address this issue and optimize the process, we have decided to suggest that the BLASTp implementation must be revised. Our revised approach would involve downloading a part of the NCBI BLAST database and executing the BLASTp searches locally. This adjustment not only streamlines the workflow but also ensures that the pipeline can effectively handle larger datasets without overburdening external databases. Due to time constraints this is outside the constraints of the project.

Microprotein characteristics

When examining microproteins, several protein characteristics prove to be of interest, as they provide insights into the nature and behavior of these microproteins. These characteristics include:

- Hydrophobicity: Hydrophobicity is a fundamental property that plays a crucial role in protein folding. It pertains to the affinity of amino acids within a protein for water molecules and influences the protein's three-dimensional structure.
- Isoelectric point: The isoelectric point represents the pH value at which the net charge of a protein becomes zero. It provides information about the solubility of the protein under different pH conditions.
- Instability Index: The instability index is an indicator of the protein’s stability. For instance, antimicrobial peptides are considered stable when their instability index values are less than 40.
- Mass over charge: mass over charge, charge fixed for the weight of the protein.
- Length: The length of the protein in amino acids.
- Molecular weight: Molecular weight represents the total weight of the microprotein in Daltons (Da).

- Charge: the charge of the protein by a specific pH, interchangeable with hydrophobicity.

To assess and calculate these critical characteristics of microproteins, we have utilized the Peptides package in R (Osorio et al. 2023). This package contains functions designed to predict these characteristics based on the amino acid sequence of the microprotein. By leveraging these computational tools, we can gain an understanding of the physicochemical properties of microproteins, facilitating their characterization and functional analysis.

Final overview pipeline

The constructed pipeline serves as a powerful framework for the simultaneous execution of multiple analytical tools, allowing us to computationally analyze microproteins. This pipeline efficiently combines OmegaFold, DeepTMHMM, SignalP, netMHCpan, the Peptides package for characteristics prediction, and IUPred3, and a SLiM search. A schematic overview of the pipeline is depicted in figure 4.

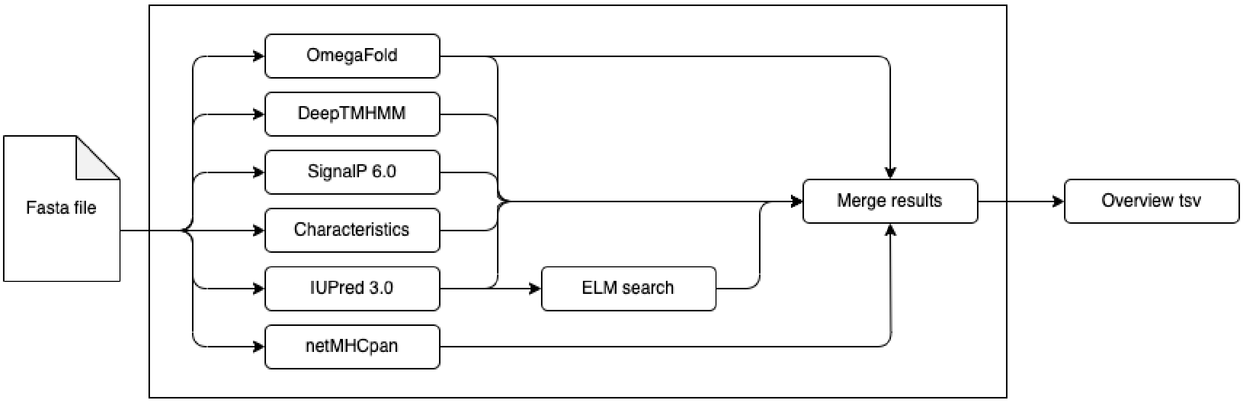


Figure 4: Schematic overview of constructed pipeline. It needs a fasta file as input and outputs a tab separated file containing the results of all executed tools. Tools are executed parallel.

Pipeline components and output:

- Structure prediction based on OmegaFold
    - Output:
        - pLDDT score average & per residue score
        - pLDDT score > 70 is considered reliable
- Disorder prediction of the residues with IUPred 3.0
    - Output:
        - Average disorder and per residue disorder scores
        - A disorder score above 0.5 means disordered
- SLiM search
    - Output:
        - SLiMs In the whole protein and SLiMs that are only in the disordered parts of the protein
        - Sequences, elm class names and total SLiM count
- Several characteristics, like hydrophobicity and iso electric point are predicted with the peptides package in R.

    - Output:
        - Predicts and reports various characteristics, including hydrophobicity, isoelectric point, instability, mass over charge, length, molecular weight, and charge at specific pH levels.
- Prediction of MHC-I binding peptides with NetMHCpan.
    - Output:
        - Strong binders and weak binders
        - Sequences of binders and count per microprotein
- Prediction of a secretion signal with SignalP 6.0.
    - Output:
        - Presence of secretion signal
- Predict presence of trans membrane domains with DeepTMHMM
    - Output:
        - Options that are predicted: transmembrane (TM), globular (Glob), transmembrane + secretion peptide (TM+SP) and, secretion peptide (SP).

This pipeline efficiently generates comprehensive microprotein profiles, including structural, functional, physicochemical, and localization characteristics. The output is provided in a tab-separated file, which can be used for further visualization and prioritization of proteins for additional research. Intermediate files are also saved, allowing for in-depth data inspection and analysis.

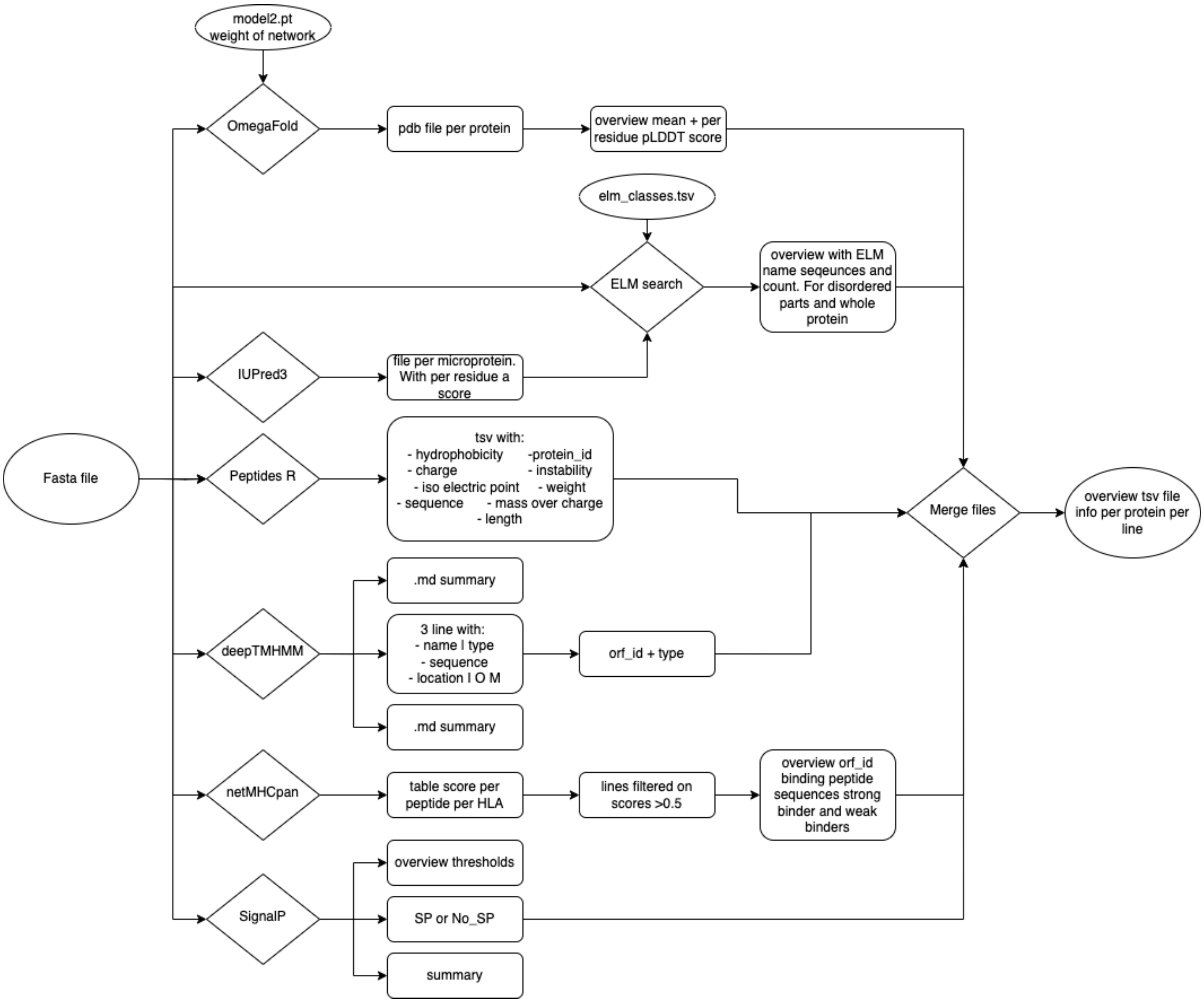

model2.pt weight of network
OmegaFold
pdb file per protein
overview mean + per residue pLDDT score
elm_classes.tsv
ELM search
overview with ELM name seqeunces and count. For disordered parts and whole protein
IUPred3
file per microprotein. With per residue a score
Fasta file
Peptides R
tsv with:
- hydrophobicity
- charge
- iso electric point
- sequence
-protein_id
- instability
- weight
- mass over charge
- length
Merge files
overview tsv file info per protein per line
.md summary
deepTMHMM
3 line with:
- name I type
- sequence
- location I O M
orf_id + type
.md summary
netMHCpan
table score per peptide per HLA
lines filtered on scores >0.5
overview orf_id binding peptide sequences strong binder and weak binders
overview thresholds
SignalP
SP or No_SP
summary

Figure 5: Flowchart of pipeline. Depicted in squares are the tools / scripts. Depicted in ovals are the input files and the output file. Depicted in rounded squares are the intermediate files.

Limitations of the pipeline

Due to server load and time restrictions on external servers, there is a maximum limit on the number of proteins that can be analyzed at once, particularly for the tool that submits data externally (e.g., DeepTMHMM). When this limit is exceeded, tools may not start or may halt prematurely. Both excessively short and long proteins can pose challenges for the pipeline. IUPred 3.0 requires a minimum protein length of 19 amino acids for calculations. Longer proteins require significantly more time for processing in the pipeline. Longer proteins require more processing time, so analyzing large datasets with lengthy proteins should be done with caution, it is advisable to analyze them in smaller batches. While the tools provide predictions, the thresholds and values per prediction need to be accessed from the intermediate files. Additionally, experimental validation is essential to confirm the accuracy of these predictions. In summary, the pipeline offers a robust and comprehensive solution for microprotein analysis, encompassing diverse characteristics and properties. However, it is important to be mindful of its limitations to ensure optimal use and accurate interpretation of the results.

Resources needed.

Resources needed for the pipeline are dependent on the number of proteins analyzed at once and on their size. A lot of small proteins will take the same amount of time as fewer long proteins. The recommended resources are based on amino acids rather than number of microproteins. It is recommended to split the input FASTA file in several smaller FASTA files or give the pipeline more resources if the tools with the provided computed resources fail.

DeepTMHMM is the bottleneck of the pipeline, as it can only efficiently analyze roughly 225,000 to 250,000 amino acids at once. It might handle slightly more but exceeding around 450,000 amino acids is likely to lead to failure. Start time is also a consideration, as a busier cluster may require more time until tool execution. All other tools are used on our own cluster and can be scaled up for processing any number of amino acids. However, it is important to note that netMHCpan may require a significant amount of time when processing a single, large Fasta file. For instance, analyzing 250,000 amino acids with netMHCpan takes approximately 16 hours.

Table 7: overview of resources used based on a fasta file with 250000 amino acids.

Tool	*Time*	*Mem*	*Tmp space*	*GPU / CPU*
OmegaFold	5:00:00	10 Gb	-	GPU
NetMHCpan	16:00:00	500 Mb	250 Mb	CPU
IUPred	1:00:00	1 Gb	-	CPU
SignalP	00:30:00	3 Gb	-	CPU
DeepTMHMM	1:00:00	1 Gb	-	CPU
peptides	00:05:00	10 Mb	10 Mb	CPU
SLiM search	00:15:00	1 Gb	500 Mb	CPU

Results

As first use case of the constructed pipeline we used a dataset of 15 neuroblastoma ribo-seq samples. These samples were mapped to a RNA-seq constructed transcriptome and information about the ORFs was collected. Mapping these ORFs against the known proteome gave us 2464 microproteins (19-150aa) that were non-canonical in neuroblastoma. As a control dataset and as a comparison, a set of 3120 canonical microproteins (19-150aa) from the same dataset in neuroblastoma was used.

We hypothesized that there would be differences between canonical and non-canonical proteins. The non-canonical microproteins would be shorter in general, would have less p sites per residue, are more likely to be disordered, harder to predict a structure from and are suspected to have less signals and transmembrane domains present. This because of their probable disordered origin and because microproteins are evolutionary young (Sandmann et al. 2023; Wilson et al. 2017). Nevertheless, we hypothesize to find interesting candidates for further research.

Neuroblastoma canonical ORFs

For all canonical ORFs predictions were made with the pipeline. In the canonical ORF dataset there were 3120 microproteins. From these 3120 canonical ORFs there are 9 subcategories of ORF types. Most of them are in the category ORF_annotated or N_truncation which is suspected when looking at canonical microproteins (table 9). From the 3120 canonical ORFs 1581 have a OmegaFold score above 70 which is reliable, 205 are predicted to have a signal peptide by Signal P, 469 are predicted to be transmembrane by DeepTMHMM, and 372 and are predicted to have a signal by DeepTMHMM (table 10). SignalP and DeepTMHMM overlap with predicting a signal for 174 of these ORFs (fig. 6). The canonical microproteins have a mean length of 101, a mean hydrophobicity of 5.48 a mean charge of 1.83, a mean of 4.8 SLiMs per disordered regions of the proteins, and a mean of 21.5 strong binders to the MHC per protein (table 8).

Table 8: Summary of Canonical Microprotein Characteristics

Characteristic	*Mean Value*	*Standard Deviation*	*Range*
length (amino acids)	101	33	19 – 150
Hydrophobicity	5.48	0.35	4.13 - 7.78
SLiMs	4.80	6.20	0 - 41
Strong Binders	21.5	10.6	0 - 60
Charge	1.83	7.2	-44.1 – 40.1

Table 9: Count of canonical ORF types

C_extension	*C_truncation*	*N_extension*	*N_truncation*	*NC_extension*	*Nested_ORF*	*ORF_annotated*	*Overl_dORF*	*Overl_dORF*
1	12	26	869	5	27	1853	88	239
0.03%	0.4%	0.8%	27.9%	0.2%	0.9%	59.4%	2.8%	7.7%

Table 10: Count of ORFs above threshold in dataset with canonical ORFs.

Tool	*Count*
OmegaFold >0.7	1581 (51%)
SignalP SP	205 (7%)
DeepTMHMM TM	469 (15%)
DeepTMHMM SP	372 (12%)

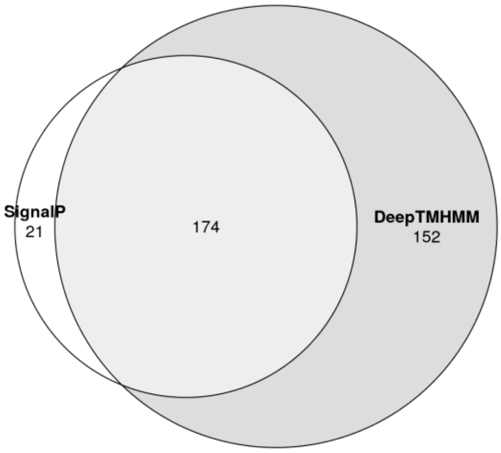


Figure 6: Venn diagram of presence of signal in canonical microproteins. Predicted by SignalP and DeepTMHMM.

Neuroblastoma non-canonical ORFs

For all non-canonical ORFs predictions were made with the pipeline. In the non-canonical ORF dataset there were 2464 microproteins. From these 2464 non-canonical ORFs there are 4 subcategories of ORF types. 232 lncORFs, 594 uORFs, 114 dORFs, and 1493 novel ORFs (table 12). From the 2464 non-canonical ORFs 781 have a OmegaFold score above 70 which is reliable, 56 are predicted to have a signal peptide by Signal P, 106 are predicted to be transmembrane by DeepTMHMM, and 392 and are predicted to have a signal by DeepTMHMM (table 13). SignalP and DeepTMHMM overlap with predicting a signal for 49 of these ORFs (fig. 7). The non-canonical microproteins have a mean length of 75, a mean hydrophobicity of 5.25, a mean charge of 1.83, a mean of 5 SLiMs per disordered regions of the proteins, and a mean of 15 strong binders to the MHC per protein (table 11).

Table 11: Summary of non-canonical Microprotein Characteristics

Characteristic	*Mean Value*	*Standard Deviation*	*Range*
length (amino acids)	74.5	37.0	19 - 150
Hydrophobicity	5.25	0.36	4.25 – 7.08
SLiMs	4.80	6.20	0 - 33
Charge	1.83	6.03	-42.94 – 44.37
Strong Binders	15.27	10.65	0 - 53

Table 12: count of non-canonical ORF types

lncORF	*uORF*	*dORF*	*novel*
232	594	114	1493
9.5%	24,4%	4,7%	61,2%

Table 13: Count of ORFs above threshold in dataset with non-canonical ORFs.

Tool	*Count*
OmegaFold >0.7	781 (30%)
SignalP SP	56 (2%)
DeepTMHMM TM	106 (4%)
DeepTMHMM SP	392 (15%)

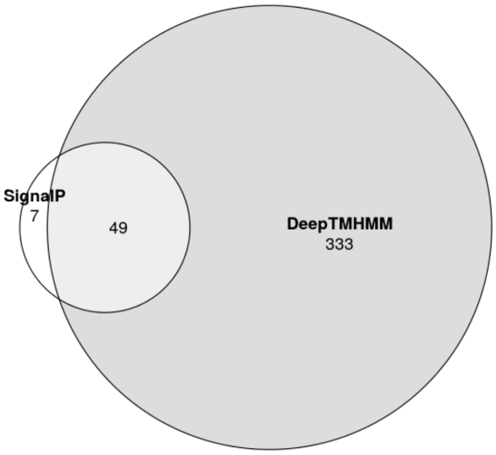


Figure 7: Venn diagram of presence of signal in non-canonical microproteins. Predicted by SignalP and DeepTMHMM

Comparison of canonical and non-canonical ORFs

Similarities

Most characteristics of microproteins are the same whether we look at the canonical dataset or the non-canonical dataset. The mean IUPred3.0 scores were similar (fig 8. B) and the hydrophobicity was similar (fig 8. C), Similar structural motif and functional domain counts indicate common features that may play essential roles in microprotein function across both categories.

Differences

There are differences in length (fig 8. A), p-sties per residue (fig 8. E), and OmegaFold pLDDT score (fig 8. D) between canonical and non-canonical microproteins in neuroblastoma. There was also a difference in the number of predicted Signal Peptides and Transmembrane proteins (table 10 & 13). The difference in size suggests that most canonical proteins are longer than non-canonical protein, which confirms the bias towards longer proteins which are described in literature. The difference in p-sites per residue suggests that canonical microproteins are translated more than non-canonical proteins which confirms the translation bias towards canonical proteins. The difference in OmegaFold pLDDT score suggests that canonical microprotein structure can be predicted more reliably than non-canonical proteins, which was suspected. The difference in number of signal peptides and transmembrane proteins suggest that there are less functional proteins in the non-canonical dataset than in the non-canonical dataset, which was already hypothesized.

All differences found are in line with what we hypothesized based on the present literature.

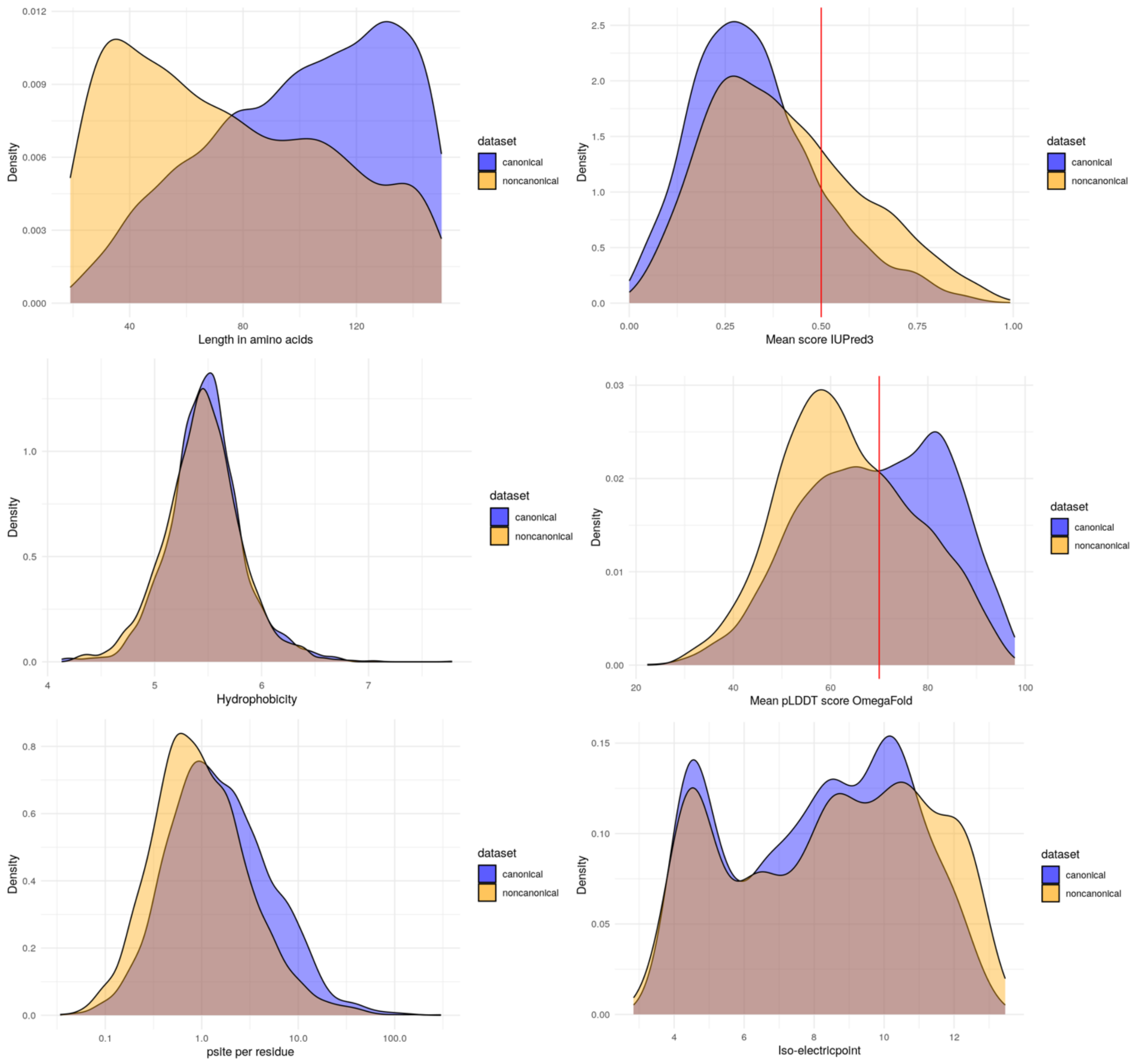


Figure 8: Density plots of canonical and non-canonical proteins of neuroblastoma. Blue is non-canonical and red is canonical. A) Density plot of the length. B) Density plot of the mean IUPred3 score red line depicts the disorder score of 0.5. C) Density plot of the hydrophobicity. D) Density plot of the mean pLDDT scores of OmegaFold, red line depicts the pLDDT score of 70. E) Density plot of the p site per residue. F) Density plot of iso-electric point.

Neuroblastoma enriched ORFs

To find ORFs in neuroblastoma that are interesting to prioritize and possibly be used for immunotherapy and/or cellular therapies. The list of all non-canonical microproteins was filtered on neuroblastoma enriched ORFs. This is done by comparing the neuroblastoma RNA sequencing data with other datasets. This gave us 147 neuroblastoma enriched transcripts and 25 of these transcripts were found in our ribo-seq data. From these 25 neuroblastoma enriched ORFs there are 3 subcategories of ORF types. There are 2 lncORFs, 11 uORFs and 12 novel ORFs (table 14). Different types of ORFs can have different transcription rates and functions, which can be interesting and important for their function. From the 25 neuroblastoma enriched ORFs 7 have a OmegaFold score above 70 which is reliable, 1 is predicted to have a signal peptide by Signal P, 1 is predicted to be transmembrane by DeepTMHMM, and 5 and predicted to have a signal by DeepTMHMM (table 15). SignalP and

DeepTMHMM overlap with predicting a signal for 1 of these ORFs (fig. 9). These ORFs are the first 11 that are extra of interest because of their potential biological function and/or place within the cell.

Table 14: Neuroblastoma enriched ORF types.

lncORF	*uORF*	*dORF*	*novel*
2	11	-	12
8%	44%	-	48%

Table 15: Interesting properties count from neuroblastoma enriched dataset, biological relevance. Total of 25 microproteins enriched for neuroblastoma.

Tool	*Count*
OmegaFold > 70	7 (28%)
SignalP SP	1 (4%)
DeepTMHMM TM	1 (4%)
DeepTMHMM SP	5 (20%)

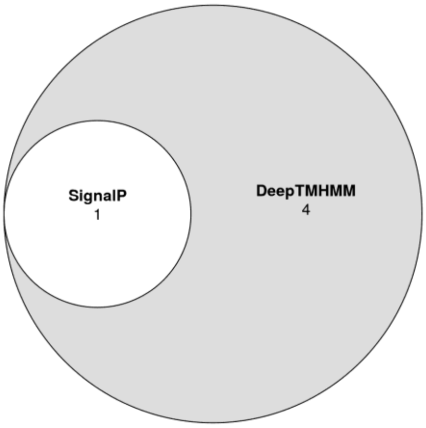


Figure 9: Venn diagram of presence of signal in neuroblastoma enriched microproteins. Predicted by SignalP and DeepTMHMM.

Interesting neuroblastoma enriched ORFs

Out of the 25 neuroblastoma enriched ORFs three ORFs were selected to be the most potential and interesting targets. They include protein target_1, target_2, and target_3**.** All three microproteins are present in 3 or more samples have a good 3D prediction and/or are predicted to have a signal peptide or be transmembrane. An overview of the properties of these three proteins can be found in table 16. We hypothesize that with further research it could be determined whether the predictions about these proteins are true and whether they are potential targets for immunotherapy.

Table 16: The interesting neuroblastoma enriched microproteins with biological relevance based on the pipeline.

	Target_1	*Target_2*	*Target_3*
Length	117	35	74
OmegaFold score	80	80	63
DeepTMHMM prediction	Transmembrane	Signal present	Signal present
Sample count	7 (47%)	3 (20%)	3 (20%)
Mean disorder score	0.08	0.25	0.5
Type ORF	uORF	uORF	uORF
BLASTp result	Hits, L-type amino acid transporter	No hits	No hits
DeepLoc result	cytoplasm	cytoplasm	Extracellular
Strong binders	35	8	22
Weak binders	73	11	45
SignalP	No signal	No signal	Secretion signal

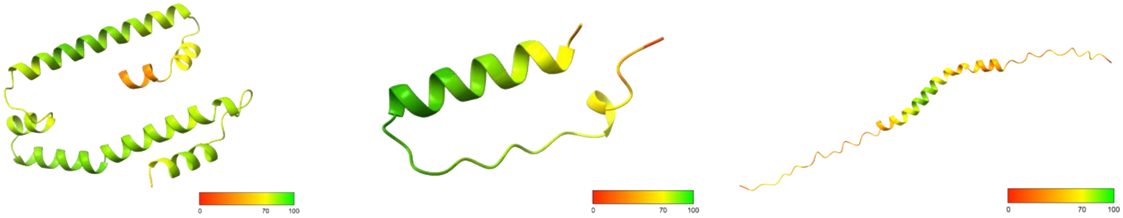


Figure 10:. A) target_1 B) target_2 C) target_3 they are all predicted by OmegaFold and visualized with chimeraX. Colors indicate the pLDDT score predicted by OmegaFold.

The first interesting example of neuroblastoma enriched microprotein is target_1 this microprotein is present in 7 out of the 15 samples (47%). It is predicted to be transmembrane and OmegaFold predicts the 3D structure with High (>80) confidence (fig 10. A) and has a lot of strong and weak binders. These properties together hint towards a functional protein which is probably presented by an HLA molecule. A second interesting example of neuroblastoma enriched microprotein is target_2 this microprotein is present in 3 out of 15 samples (20%). It is predicted to have a signal present and OmegaFold predicts a 3D structure with high (>80) confidence (fig 10. B). Predictions suggest that this microprotein has a signal and is of high confidence predicted value, this makes the protein interesting to investigate further. A third interesting example of a microprotein that is neuroblastoma enriched and looks interesting based on the predictions of the pipeline is target_3 (fig 10. C)**.** SignalP 6.0, deepTMHMM and DeepLoc 2.0 predict that this protein has a signal peptide present and is localized outside of the cell. This together with its appearance in 3 of 15 samples (20%) and substantial number of strong binders to the MHC-I makes it an interesting microprotein. It would be interesting to research whether this protein is really secreted and potentially has a function outside of the cell.

In conclusion, these microproteins are the most interesting targets based on the pipeline for further research.

Discussion

The Pipeline's Success

The pipeline proved to be a valuable tool in identifying various properties of microproteins. We were able to calculate characteristics that allowed us to prioritize specific microproteins for further investigation. We were also able to predict present signal peptides, trans membrane parts, the 3D structure, short linear motifs, disorder, and the binding of peptides to the MHC-I. Looking at all these results combined we were able to find potential functions and localizations of several of the non-canonical microproteins of interest. With this it is possible to predict biological function of non-canonical microproteins and prioritize them based on the results by what is interesting.

We successfully combined the peptides package in R, OmegaFold, IUPred 3.0, SignalP 6.0, DeepTMHMM, netMHCpan, and a SLiM search. As first use case we looked at neuroblastoma, a pediatric cancer that has low survival rates for the high-risk variant. Neuroblastoma accounts for 12 - 15% of childhood cancer deaths and it would benefit greatly from a different therapy strategy. The biological functions and potential for immunotherapy of microproteins in neuroblastoma are of interest for this. In our dataset which included 2464 non-canonical microproteins. We suggest further research for 3 candidates found based on the results of the pipeline. Candidate 1: Target_1 is interesting because the proteins is predicted to be transmembrane combined with a good OmegaFold score and a low disorder score. Further research could focus on first determining whether the protein has a crucial role in the tumor when silencing the microprotein. After that trying to confirm the presence of the microprotein in het membrane of the cell. Candidate 2: Target_2 is interesting because it has a good OmegaFold score combined with a low disorder score. Further research could focus on determining the function with silencing the ORF. Candidate 3: Target_3 is interesting because DeepTMHMM, SignalP and DeepLoc predict that it has a signal, is secreted and is extracellular. Further research could focus on determining whether this protein is secreted and whether it has a function outside of the cell.

The Pipeline’s Limitations

Length and size constraints

One notable constraint was the pipeline's runtime when applied to longer proteins. The runtime increased significantly with long proteins (> 150aa), posing a challenge for analyzing a lot of larger molecules at a time. To mitigate this, it is beneficial to run the pipeline in smaller batches. Another limitation was the minimum length of microproteins of 19 amino acids which was needed to run the pipeline, which is due to IUPred 3.0 which needs proteins to at least have this length. All microproteins shorter than that have to be discarded.

Workflow Manager

In our current workflow, when a tool within the pipeline fails due to input issues or other unforeseen circumstances, manual intervention is necessary. Which means identifying the failed tool, addressing the issue, and manually restarting the pipeline or parts of it. Additionally, we must handle the merging of results from various tools by hand, a task that can become increasingly complex as the pipeline expands and more tools are integrated. A workflow manager might be a good solution to overcome these issues. This is how a workflow manager can enhance our pipeline. Automated Error Handling: With a workflow manager in place, the handling of tool failures becomes automated. When a tool encounters an issue, the workflow manager can detect it and initiate the necessary actions for resolution. This eliminates the need for manual intervention, saving time and reducing the risk of errors introduced during manual restarts. Efficient Resource Utilization: By managing the pipeline's execution, a workflow manager can optimize resource allocation. It can ensure that computational resources are efficiently distributed among the tools, reducing the risk of resource contention or underutilization. This, in turn, enhances the overall performance and speed of the pipeline. Selective Reruns: One of the most significant advantages of a workflow manager is its ability to selectively rerun only the tools that fail, rather than restarting the entire pipeline or restarting manually. Enhanced Scalability: As our research evolves and the complexity of our pipeline grows, scalability becomes a crucial consideration. A workflow manager lays the foundation for scalability, as it can easily accommodate the addition of new tools or modules without fundamentally altering the pipeline's structure. Time and resource constraints during our study limited our ability to incorporate a workflow manager. Additionally, the adoption of a workflow manager may require a learning curve and adjustments to the existing pipeline structure.

Navigating Biases in Databases

There is still an issue in the field of microprotein research, the biases in existing databases which can result in a bias when neural networks are trained for predictions (Kleppe et al. 2021). These biases stem from the focus on well-studied proteins, which has led to extensive annotations and data for these proteins. Microproteins, however, represent a relatively uncharted territory. Consequently, they may be underrepresented or entirely absent from many databases. This lack of annotations for microproteins does not imply their non-existence but rather reflects a gap in our knowledge and exploration. It underscores the importance of conducting further studies on these microproteins. This realization should drive efforts to expand and enrich databases with data on microproteins, gradually reducing the bias against them.

Significance of SLiMs in Context

Short linear motifs (SLiMs) offer a fascinating window into the functional aspects of microproteins. These motifs provide insights into potential interactions and roles within cellular processes. However, they come with their own complexity. It is important to note that SLiMs can occur by chance, independent of any functional significance (Davey et al. 2011; Neduva and Russell 2005). Which makes it essential to avoid jumping to conclusions solely based on the presence of SLiMs (Van Roey, Gibson, and Davey 2012). A nuanced approach involves considering SLiMs in combination with other microprotein properties. For instance, identifying specific SLiMs within a certain group of microproteins, such as those with signal peptides, could be informative. These combinations of characteristics may point towards more significant functional roles, enhancing our understanding of microprotein biology.

MHC-I binding peptides and their significance

Currently our pipeline takes all peptides from 8 – 12 amino acids long, that our microproteins could potentially split in, into account. The binding of these peptides to 12 different types of HLA is predicted and all strong and weak binders are putted out. This is interesting and gives info about the possible presentation of the microproteins on the outside of cells. But with the goal of immunotherapy in mind it is important to subset the binders on whether they are already known to bind the MHC-I complex or not. We propose to add a component to the pipeline that filters all the known HLA binders. This could add an extra column with new strong and weak binders that are unique. Microproteins that have strong and weak binders predicted that are not known in immunopeptidomics data are the most interesting for prioritizing in future research.

Another important thing to note is that there are biases when looking at MHC-I binding affinity, HLA subtypes are present in parts of the population and the MHC-I complex has a preferable peptide length when assembling on the complex. First, HLA subtypes are only expressed within subtypes of the population (Wang and Claesson 2014). By selecting the 12 HLA types that are most common within the world population we try to account for this. The HLA supertypes were used for this, which were reported to cover most of the HLA-A and -B polymorphisms (Wang and Claesson 2014). Using different types of HLA allows us to predict in which population the peptides would be presented. Second, the preferable peptide length of MHC-I is 9 amino acids, netMHCpan which is trained on MHC peptide data also has a bias for 9-mers when predicting binders. But it allows for differences in length by incorporating a single alignment step which allows for insertions and/or deletions (Nielsen and Andreatta 2016). With this we try to account for the differences in HLA types within the population and the 9-mer bias of the MHC-I complex.

Future Considerations

Mass Spectrometry Data Analysis

Mass spectrometry is a powerful technique for identifying and quantifying molecules, including proteins. In the context of microproteins, mass spectrometry could potentially offer a direct method for confirming the presence of these microproteins in biological samples. The size of microproteins presents unique difficulties, especially with mass spectrometry. Cleavage Limitations: Mass spectrometry relies on cleaving proteins into smaller fragments for analysis. However, the size of microproteins can make them resistant to cleavage, leading to incomplete or inconclusive results. This poses a significant challenge in obtaining comprehensive data about the structural properties of these molecules. Smaller Size limitations: Microproteins, typically less than 100 amino acids in length, fall below the lower limit for conventional mass spectrometry analysis. This size constraint can make it inherently difficult to obtain reliable mass spectrometry data, as the technique is optimized for larger proteins. New methodologies or adaptations of mass spectrometry techniques may be needed to provide more definitive results regarding microprotein stability. Finding microproteins in mass spectrometry data would be a direct indication that a protein is present in cells and that a protein is stable. This is crucial when a protein has a function. In future research, when more time becomes available, exploring mass spectrometry data could be immensely valuable. It could help us validate the existence of specific microproteins in various biological contexts, shedding light on their expression, stability, and potential functions.

Immunopeptidomics

Immunopeptidomics is a field that focuses on the identification and characterization of peptides presented on the cell surface by the major histocompatibility complex (MHC) (Chong, Coukos, and Bassani-Sternberg 2022). These peptides play a crucial role in immune recognition. Investigating microproteins in the context of immunopeptidomics could reveal whether these small proteins are involved in immune responses or other cellular processes. While our study did not delve into this area due to time constraints, it is an intriguing avenue to explore. Looking at peptides from microproteins present in immunopeptidomics data could provide insights into their interactions with the immune system and their potential significance in health and disease.

Protein-Protein Interactions

Exploring protein-protein interactions involving microproteins is a complex challenge. Many existing tools are tailored to larger, well-characterized proteins, making them less effective for studying microproteins. Our study briefly touched on this topic but did not pursue it extensively. Future research in this area could involve developing specialized tools or leveraging advancements in the field to better understand how microproteins interact with other cellular components. This knowledge could unveil the roles of microproteins in intricate biological networks, potentially leading to the discovery of novel therapeutic targets.

Weighted Gene Correlation Network Analysis

A first step when looking at protein-protein interactions could be Weighted Gene Correlation Network Analysis (WGCNA)(Langfelder and Horvath 2008). WGCNA looks into the relationships between genes, aiming to uncover whether a gene's expression is linked with the expression patterns of other genes. This approach is far more nuanced than merely assessing individual gene expression levels in isolation. WGCNA provides valuable insights into the cooperative behavior of genes within a biological system. It allows us to discern whether a particular gene tends to be co-expressed with a set of other genes in a non-random manner. This non-random co-expression is often a key indicator that the genes involved share common regulatory mechanisms, participate in the same biological pathways, or even interact directly. While WGCNA is primarily based on RNA-seq data, it holds great promise when combined with our experimental results. This combination can lead to better interpretations of our data, potentially uncovering associations and biological insights.

Alternative Homology Prediction Tools

Homology prediction tools, such as BLAST and HMMER, are important for identifying evolutionary relationships between proteins. However, as we encountered in our study, using these tools for microproteins, particularly those that are evolutionarily young or funique, can be challenging. Exploring alternative homology prediction methods tailored to the specific characteristics of microproteins is a promising avenue. These tools could enable researchers to uncover evolutionary connections, functional insights, and potential conserved domains in microproteins that might have been missed by more traditional approaches.

FoldSeek for divergent evolved microproteins

To not only look at conservations but also at divergent evolved microproteins. FoldSeek could be added to our pipeline, with FoldSeek the visual representations are mapped to known 3D structures. This overview will provide an understanding of functions that a protein may have based on structure rather than conservation. The tool that is developed for this is FoldSeek, it is relatively new and uses the 3D structure to find proteins that have a similar structure (van Kempen et al. 2023). Due to time constraints, we did not add this to our pipeline, but it would be nice to add in the future.

In summary, our study laid the groundwork for understanding microproteins and their properties. However, the complexities and nuances of these microproteins, coupled with limitations in time and resources, left some avenues unexplored. These areas, including mass spectrometry data analysis, immunopeptidomics, protein-protein interactions, and alternative homology prediction tools, hold immense potential for future research. As the field advances and as more data becomes available, revisiting these avenues could uncover insights into microprotein biology, further enriching our understanding.

Bibliography

Abbott, Maura, and Yelena Ustoyev. 2019. 'Cancer and the Immune System: The History and Background of Immunotherapy'. *Seminars in Oncology Nursing* 35(5): 150923.

Ahdritz, Gustaf et al. 2022. 'OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization'. *bioRxiv*: 2022.11.20.517210.

Altschul, S. F. et al. 1997. 'Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs'. *Nucleic Acids Research* 25(17): 3389–3402.

Armenteros, Jose Juan Almagro et al. 2019. 'Detecting Sequence Signals in Targeting Peptides Using Deep Learning'. *Life Science Alliance* 2(5). https://www.life-science-alliance.org/content/2/5/e201900429 (September 19, 2023).

Aubel, Margaux, Lars Eicholt, and Erich Bornberg-Bauer. 2023. 'Assessing Structure and Disorder Prediction Tools for *de Novo* Emerged Proteins in the Age of Machine Learning'. https://f1000research.com/articles/12-347 (October 2, 2023).

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. 'The MEME Suite'. *Nucleic Acids Research* 43(W1): W39–49.

Blees, Andreas et al. 2017. 'Structure of the Human MHC-I Peptide-Loading Complex'. *Nature* 551(7681): 525–28.

Chen, Hanbo. 2022. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. https://cran.r-project.org/web/packages/VennDiagram/index.html (October 25, 2023).

Chong, Chloe, George Coukos, and Michal Bassani-Sternberg. 2022. 'Identification of Tumor Antigens with Immunopeptidomics'. *Nature Biotechnology* 40(2): 175–88.

Chothani, Sonia P. et al. 2022. 'A High-Resolution Map of Human RNA Translation'. *Molecular Cell* 82(15): 2885-2899.e8.

Chowdhury, Ratul et al. 2022. 'Single-Sequence Protein Structure Prediction Using a Language Model and Deep Learning'. *Nature Biotechnology* 40(11): 1617–23.

Conte, Alessio Del et al. 'Critical Assessment of Protein Intrinsic Disorder Prediction (CAID) - Results of Round 2'. *Proteins: Structure, Function, and Bioinformatics* n/a(n/a). http://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26582 (October 2, 2023).

Davey, Norman E. et al. 2011. 'Attributes of Short Linear Motifs'. *Molecular BioSystems* 8(1): 268–81.

Dinger, Marcel E., Ken C. Pang, Tim R. Mercer, and John S. Mattick. 2008. 'Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities'. *PLOS Computational Biology* 4(11): e1000176.

Dobin, Alexander et al. 2013. ‘STAR: Ultrafast Universal RNA-Seq Aligner’. *Bioinformatics (Oxford, England)* 29(1): 15–21.

DUAN, YOU et al. 2021. ‘A Systematic Evaluation of Bioinformatics Tools for Identification of Long Noncoding RNAs’. *RNA* 27(1): 80–98.

Edwards, Richard J., and Nicolas Palopoli. 2015. ‘Computational Prediction of Short Linear Motifs from Protein Sequences’. In *Computational Peptidology*, Methods in Molecular Biology, eds. Peng Zhou and Jian Huang. New York, NY: Springer, 89–141. https://doi.org/10.1007/978-1-4939-2285-7_6 (September 22, 2023).

Erdős, Gábor, Mátyás Pajkos, and Zsuzsanna Dosztányi. 2021. ‘IUPred3: Prediction of Protein Disorder Enhanced with Unambiguous Experimental Annotation and Visualization of Evolutionary Conservation’. *Nucleic Acids Research* 49(W1): W297–303.

Filbin, Mariella, and Michelle Monje. 2019. ‘Developmental Origins and Emerging Therapeutic Opportunities for Childhood Cancer’. *Nature Medicine* 25(3): 367–76.

Gabler, Felix et al. 2020. ‘Protein Sequence Analysis Using the MPI Bioinformatics Toolkit’. *Current Protocols in Bioinformatics* 72(1): e108.

Gíslason, Magnús Halldór, Henrik Nielsen, José Juan Almagro Armenteros, and Alexander Rosenberg Johansen. 2021. ‘Prediction of GPI-Anchored Proteins with Pointer Neural Networks’. *Current Research in Biotechnology* 3: 6–13.

Gutierrez, Santiago et al. 2022. *MembraneFold: Visualising Transmembrane Protein Structure and Topology*. Bioinformatics. preprint. http://biorxiv.org/lookup/doi/10.1101/2022.12.06.518085 (September 19, 2023).

Hallgren, Jeppe et al. 2022. *DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks*. Bioinformatics. preprint. http://biorxiv.org/lookup/doi/10.1101/2022.04.08.487609 (September 19, 2023).

Hanson, Jack, Kuldip K. Paliwal, Thomas Litfin, and Yaoqi Zhou. 2019. ‘SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning’. *Genomics, Proteomics & Bioinformatics* 17(6): 645–56.

Harris, Charles R. et al. 2020. ‘Array Programming with NumPy’. *Nature* 585(7825): 357–62.

Hassel, Keira R., Omar Brito-Estrada, and Catherine A. Makarewich. 2023. ‘Microproteins: Overlooked Regulators of Physiology and Disease’. *iScience* 26(6). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10199267/ (October 28, 2023).

Heesch, Sebastiaan van et al. 2019. ‘The Translational Landscape of the Human Heart’. *Cell* 178(1): 242-260.e29.

Hu, Gang et al. 2021. ‘FlDPnn: Accurate Intrinsic Disorder Prediction with Putative Propensities of Disorder Functions’. *Nature Communications* 12(1).

/pmc/articles/PMC8295265/ /pmc/articles/PMC8295265/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8295265/.

Ingolia, Nicholas T., Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. 2009. ‘Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling’. *Science* 324(5924): 218–23.

Käll, Lukas, Anders Krogh, and Erik L. L Sonnhammer. 2004. ‘A Combined Transmembrane Topology and Signal Peptide Prediction Method’. *Journal of Molecular Biology* 338(5): 1027–36.

van Kempen, Michel et al. 2023. ‘Fast and Accurate Protein Structure Search with Foldseek’. *Nature Biotechnology*: 1–4.

Kleppe, Andreas et al. 2021. ‘Designing Deep Learning Studies in Cancer Diagnostics’. *Nature Reviews Cancer* 21(3): 199–211.

Kumar, Manjeet et al. 2022. ‘The Eukaryotic Linear Motif Resource: 2022 Release’. *Nucleic Acids Research* 50(D1): D497–508.

Langfelder, Peter, and Steve Horvath. 2008. ‘WGCNA: An R Package for Weighted Correlation Network Analysis’. *BMC bioinformatics* 9: 559.

Langmead, Ben, and Steven L Salzberg. 2012. ‘Fast Gapped-Read Alignment with Bowtie 2’. *Nature methods* 9(4): 357–59.

Lu, Hui-Chun, Arianna Fornili, and Franca Fraternali. 2013. ‘Protein–Protein Interaction Networks Studies and Importance of 3D Structure Knowledge’. *Expert Review of Proteomics* 10(6): 511–20.

magrittr), Stefan Milton Bache (Original author and creator of, Hadley Wickham, Lionel Henry, and RStudio. 2022. *Magrittr: A Forward-Pipe Operator for R*. https://cran.r-project.org/web/packages/magrittr/index.html (October 25, 2023).

Martin, Marcel. 2011. ‘Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads’. *EMBnet.journal* 17(1): 10–12.

Matthay, Katherine K. et al. 2016. ‘Neuroblastoma’. *Nature Reviews Disease Primers* 2(1): 1–21.

Meng, Elaine C. et al. ‘UCSF ChimeraX: Tools for Structure Building and Analysis’. *Protein Science* n/a(n/a): e4792.

Merino-Valverde, Iñaki, Emanuela Greco, and María Abad. 2020. ‘The Microproteome of Cancer: From Invisibility to Relevance’. *Experimental Cell Research* 392(1): 111997.

Mooney, Catherine, Gianluca Pollastri, Denis C. Shields, and Niall J. Haslam. 2012. ‘Prediction of Short Linear Protein Binding Regions’. *Journal of Molecular Biology* 415(1): 193–204.

Necci, Marco et al. 2021. ‘Critical Assessment of Protein Intrinsic Disorder Prediction’. *Nature Methods* 18(5): 472–81.

Neduva, Victor, and Robert B. Russell. 2005. ‘Linear Motifs: Evolutionary Interaction Switches’. *FEBS Letters* 579(15): 3342–45.

‘Neuroblastoom’. *Zorg*. https://zorg.prinsesmaximacentrum.nl/nl/diagnose/neuroblastoom (September 17, 2023).

Nielsen, Morten, and Massimo Andreatta. 2016. ‘NetMHCpan-3.0; Improved Prediction of Binding to MHC Class I Molecules Integrating Information from Multiple Receptor and Peptide Length Datasets’. *Genome Medicine* 8: 33.

Osorio, Daniel et al. 2023. *Peptides: Calculate Indices and Theoretical Physicochemical Properties of Protein Sequences*. https://cran.r-project.org/web/packages/Peptides/index.html (October 25, 2023).

Palopoli, Nicolas, Kieren T. Lythgow, and Richard J. Edwards. 2015. ‘QSLiMFinder: Improved Short Linear Motif Prediction Using Specific Query Protein Data’. *Bioinformatics* 31(14): 2284–93.

Pertea, Mihaela et al. 2015. ‘StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads’. *Nature biotechnology* 33(3): 290–95.

Pettersen, Eric F. et al. 2021. ‘UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers’. *Protein Science : A Publication of the Protein Society* 30(1): 70–82.

Potter, Simon C et al. 2018. ‘HMMER Web Server: 2018 Update’. *Nucleic Acids Research* 46(W1): W200–204.

Prensner, John R. et al. 2021. ‘Noncanonical Open Reading Frames Encode Functional Proteins Essential for Cancer Cell Survival’. *Nature Biotechnology* 39(6): 697–704.

Prytuliak, Roman, Michael Volkmer, Markus Meier, and Bianca H. Habermann. 2017. ‘HH-MOTiF: De Novo Detection of Short Linear Motifs in Proteins by Hidden Markov Model Comparisons’. *Nucleic Acids Research* 45(Web Server issue): W470–77.

Quaglia, Federica et al. 2022. ‘DisProt in 2022: Improved Quality and Accessibility of Protein Intrinsic Disorder Annotation’. *Nucleic Acids Research* 50(D1): D480–87.

Reynisson, Birkir et al. 2020. ‘NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data’. *Nucleic Acids Research* 48(W1): W449.

‘SamCC-Turbo’. https://bio.tools/samcc-turbo (September 18, 2023).

Sandmann, Clara-L. et al. 2023. ‘Evolutionary Origins and Interactomes of Human, Young Microproteins and Small Peptides Translated from Short Open Reading Frames’. *Molecular Cell* 83(6): 994-1011.e18.

Schlesinger, Dörte, and Simon J. Elsässer. 2022. 'Revisiting SORFs: Overcoming Challenges to Identify and Characterize Functional Microproteins'. *The FEBS Journal* 289(1): 53–74.

Schmitz, Jonathan F., Kristian K. Ullrich, and Erich Bornberg-Bauer. 2018. 'Incipient de Novo Genes Can Evolve from Frozen Accidents That Escaped Rapid Transcript Turnover'. *Nature Ecology & Evolution* 2(10): 1626–32.

Skolnick, Jeffrey, Mu Gao, Hongyi Zhou, and Suresh Singh. 2021. 'AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function'. *Journal of chemical information and modeling* 61(10): 4827–31.

Söding, Johannes, Andreas Biegert, and Andrei N. Lupas. 2005. 'The HHpred Interactive Server for Protein Homology Detection and Structure Prediction'. *Nucleic Acids Research* 33(Web Server issue): W244–48.

Steinegger, Martin et al. 2019. 'HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation'. *BMC Bioinformatics* 20(1): 473.

Teufel, Felix et al. 2022. 'SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models'. *Nature Biotechnology* 40(7): 1023–25.

Thumuluri, Vineet et al. 2022. 'DeepLoc 2.0: Multi-Label Subcellular Localization Prediction Using Protein Language Models'. *Nucleic Acids Research* 50(W1): W228–34.

Van Roey, Kim et al. 2014. 'Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation'. *Chemical Reviews* 114(13): 6733–78.

Van Roey, Kim, Toby J Gibson, and Norman E Davey. 2012. 'Motif Switches: Decision-Making in Cell Regulation'. *Current Opinion in Structural Biology* 22(3): 378–85.

Verkuil, Robert et al. 2022. *Language Models Generalize beyond Natural Proteins*. Synthetic Biology. preprint. http://biorxiv.org/lookup/doi/10.1101/2022.12.21.521521 (September 18, 2023).

Wang, Mingjun, and Mogens H. Claesson. 2014. 'Classification of Human Leukocyte Antigen (HLA) Supertypes'. *Immunoinformatics* 1184: 309–17.

Webb, Benjamin, and Andrej Sali. 2016. 'Comparative Protein Structure Modeling Using MODELLER'. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 54: 5.6.1-5.6.37.

Wickham, Hadley, Romain François, et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://cran.r-project.org/web/packages/dplyr/index.html (October 25, 2023).

Wickham, Hadley, Winston Chang, et al. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://cran.r-project.org/web/packages/ggplot2/index.html (October 25, 2023).

Wickham, Hadley, Davis Vaughan, et al. 2023. *Tidyr: Tidy Messy Data*. https://cran.r-project.org/web/packages/tidyr/index.html (October 25, 2023).

Wickham, Hadley, and RStudio. 2022. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://cran.r-project.org/web/packages/stringr/index.html (October 25, 2023).

Wienke, Judith et al. 2021. 'The Immune Landscape of Neuroblastoma: Challenges and Opportunities for Novel Therapeutic Strategies in Pediatric Oncology'. *European Journal of Cancer* 144: 123–50.

Wilson, Benjamin A., Scott G. Foy, Rafik Neme, and Joanna Masel. 2017. 'Young Genes Are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth'. *Nature ecology & evolution* 1(6): 0146.

Withanage, Miyuraj Harishchandra Hikkaduwa, Hanquan Liang, and Erliang Zeng. 2022. 'RNA-Seq Experiment and Data Analysis'. *Methods in Molecular Biology (Clifton, N.J.)* 2418: 405–24.

Wright, Bradley W., Zixin Yi, Jonathan S. Weissman, and Jin Chen. 2022. 'The Dark Proteome: Translation from Noncanonical Open Reading Frames'. *Trends in Cell Biology* 32(3): 243–58.

Wu, Ruidong et al. 2022. *High-Resolution* de Novo *Structure Prediction from Primary Sequence*. Bioinformatics. preprint. http://biorxiv.org/lookup/doi/10.1101/2022.07.21.500999 (September 19, 2023).

Xie, Chen et al. 'A de Novo Evolved Gene in the House Mouse Regulates Female Pregnancy Cycles'. *eLife* 8: e44392.

Xu, Wenli et al. 2020. 'Ribosome Profiling Analysis Identified a KRAS-Interacting Microprotein That Represses Oncogenic Signaling in Hepatocellular Carcinoma Cells'.

Zhang, Yuanyuan, and Zemin Zhang. 2020. 'The History and Advances in Cancer Immunotherapy: Understanding the Characteristics of Tumor-Infiltrating Immune Cells and Their Therapeutic Implications'. *Cellular and Molecular Immunology* 17(8): 807–21.