UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Technology

Klavs Jermakovs

In vitro and *in silico* epitope-paratope mapping

Master's Thesis (30 ECTS)

Bioengineering Curriculum

Supervisors:

Prof. Tuomas Knowles

PhD. Erik Abner

MSci. Alekszej Morgunov

Tartu 2022

In vitro and *in silico* epitope-paratope mapping

Abstract:

The predictability of antibody-antigen binding is a prerequisite for understanding how epitope mutations lead to immune escape through antigenic drift and for improved *in silico* vaccine design. The permissible sequence variation in an epitope, which could still be recognised by a defined paratope, is limited by the need to conserve specificity and affinity of the interaction. This imposes constraints on the structural and physicochemical properties of the epitope sequence landscape.

In this work, we investigate the performance of commonly used deep learning-based protein sequence representation models in grouping related epitope sequences together. We establish a comparative baseline through pairwise sequence alignment and demonstrate that the performance of existing methods falls short of this baseline, with all methods suffering from sequence length induced bias. This highlights the need for further task-specific method development and model fine-tuning, focusing specifically on short peptide representation learning, to achieve viable performance for research applications. Additionally, we develop and implement a refined biopanning methodology which relies on next-generation sequencing to acquire a large-scale mimotope dataset against three monoclonal antibodies. We demonstrate the importance of standardisation and specific controls in acquiring unbiased high-quality results, and describe a previously unknown peptide motif recognised by an anti-influenza A monoclonal antibody.

This work lays the foundations for further high-throughput data acquisition and computational method development for addressing the challenge of predicting epitope-paratope interaction specificity.

Keywords:

Protein Representations, Deep Learning, Phage Display, Peptides and Proteins, Next-Generation Sequencing

CERCS: P310 Proteins, enzymology; P176 Artificial Intelligence; B110 Bioinformatics, medical informatics, biomathematics, biometrics

Epitoop-paratoop seondumiste kaardistamine *in vitro* ja *in silico* metoodikatel

Lühikokkuvõte:

Antikeha-antigeen seondumiste omaduste mõistmine on eeltingimus ennustamaks valguepitoopide mutatsioonide põhjustatud immuunsignaalide nõrgenemist ning tõhustamaks vaktsiinide arendamist arvutipõhiste meetoditega. Valguepitoopide järjestuste varieeruvus on aga piiratud neid seonduvate paratoopide seondumisvõimest, mistõttu tuleb epitoobil seondumisvõime hoidmiseks säilitada oma spetsiifilisus ja affiinsus paratoobi suhtes. Seega on epitoopide järjestuste muutuvus steeriliste ja füsiokeemiliste omaduste tõttu piiratud.

Käesoleva lõputöö raames uurisime me sagedasti kasutatavate süvaõppepõhiste meetodite toimivust valgujärjestuste kujutamismudelite epitoobijärjestuste rühmitamisel. Varasemalt avalikustatud epitoobijärjestuste andmestike paarikaupa joondamise kaudu lõime me referentsi, mille abil me demonstreerisime, kuidas olemasolevate metoodikate võimekus ei küündi eksperimentaalsete andmete puhul võrreldava tasemeni, kannatades valgujärjestuste pikkusest tingitud nihke all. Need tulemused rõhutavad edasiste ülesandepõhiste meetodite arendamise ning mudelite peenhäälestuse vajalikkust. Saavutamaks teaduslikes rakendustes vajalik jõudlus, tuleks edaspidi süviti keskenduda lühikeste peptiidide esituse õppimisele. Ühtlasi töötasime me välja ja rakendasime biopaneerimise metoodikat, mis tugineb järgmise põlvkonna DNA sekveneerimisel, luues seeläbi kolme erineva monoklonaalse antikehaga reageeriva suuremahulise mimotoobiandmekogu. Meie lähenemine näitlikustas proovide standardiseerimise ja spetsiifiliste kontrollide olemasolu tähtsust erapooletute kvaliteetsete tulemuste saavutamisel. Biopaneerimise tulemuste põhjal kirjeldame ka uut valgumotiivi, mis seondub A-tüüpi gripiviiruse-vastase monoklonaalse antikehaga.

See töö paneb aluse edasistele suure läbilaskevõimega andmete kogumisele ja arvutusmeetodite arendamisele, ennustamaks tõhusamini valguepitoopide ja -paratoopide seondumisi.

Võtmesõnad:

valkude esituse õpe, süvaõpe, faagi kuvamine, peptiidid ja valgud, järgmise põlvkonna sekveneerimine

CERCS: P310 Proteiinid, ensümoloogia; P176 Tehisintellekt; B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

TABLE OF CONTENTS

TERMS, ABBREVIATIONS AND NOTATIONS

BDB - Biopanning Data Bank

BSA - Bovine Serum Albumin

DNA - Deoxyribonucleic Acid

dsDNA - Double Stranded Deoxyribonucleic Acid

ELISA - Enzyme-linked Immunoassay

IPTG - Isopropyl β-D-1-thiogalactopyranoside

KNN - k-Nearest Neighbours

LB - Luria Broth

mAbs - Monoclonal Antibodies

MLM - Masked Language Modelling

MOI - Multiplicity Of Infection

MSA - Multiple Sequence Alignment

NEB - New England Biolabs

NLP - Natural Language Processing

OD_{600} - Optical Density at wavelength of 600 nm

PCA - Principal Component Analysis

PCR - Polymerase Chain Reaction

RNA - Ribonucleic Acid

ssDNA - Single Stranded Deoxyribonucleic Acid

ssRNA - Single Stranded Ribonucleic Acid

TBS - Tris-Buffered Saline

TBST - Mixture of Tris-Buffered Saline and Tween-20

Tet - Tetracycline

t-SNE - t-Distributed Stochastic Neighbor Embedding

UMAP - Uniform Manifold Approximation and Projection

UV - Ultraviolet

X-gal - 5-Bromo-4-Chloro-3-Indolyl β-D-Galactopyranoside

INTRODUCTION

Antibodies are a group of highly specific naturally occurring proteins that help the organism fight against foreign pathogens by binding their proteins - antigens. Binding between antigen and antibody is mediated by structural and physiochemical properties occurring at the respective interfaces of paratope and epitope. High-throughput experimental characterisation of the antigen epitope sequence diversity still bound by the antibody can be studied by shotgun mutagenesis (Davidson and Doranz, 2014). However, such a method is constrained by a laborious and time-consuming experimental process that allows studying only a tiny fraction of all epitope mutational possibilities. Therefore, another approach is to use highly diverse peptide libraries that mimic the epitope. Phage display technology allows for screening large diversity of random peptides against antibody and enrich a set of binding peptides that together form a mimoset bound by the antibody. Commonly used Sanger sequencing in the phage display experiments limits the amount of binding sequences obtained from the assay. Therefore, the public mimoset libraries currently contain only 20-100 peptide sequences per target thus failing to capture the full sequence diversity of binding peptides (He et al., 2015). Our work will combine the sequencing capacity of NGS with high diversity peptide phage display biopanning to obtain large-scale mimotope data against several monoclonal antibodies at once. A handful of previous studies have combined NGS with phage display panning against monoclonal antibodies (Tarnovitski et al., 2006; Hurwitz et al., 2017) but have not addressed the issue of amplification bias; our work will include additional control techniques and establish an experimental pipeline to obtain unbiased large-scale data of antibody binding peptides.

Separating binding and non-binding proteins *in silico* currently is a developing field aimed at assisting time-consuming experimental methods by identifying potential peptides that exhibit the desired function. Deep learning sequence representation methods have shown promise in protein-protein interaction prediction (Kimothi et al., 2019). These methods represent protein sequence as a numerical vector that contains some level of information characterising its biological and structural properties (Heinzinger et al., 2019). However, such representation methods have not yet been applied in the context of epitope-paratope mapping. Our work will implement various protein sequence representation methods and evaluate their performance on both publicly available and the experimental data generated in this work.

1 Literature Overview

1.1 Phage Display

1.1.1 Overview of the Technique

Phage display is an *in vitro* technique first described in 1985 (Smith, 1985). This method physically links genotype and phenotype by fusing phage coat protein with foreign protein. As a result the foreign protein is displayed on the surface of the modified phage particle, while genetic information is retained in the phage genome. The method allows displaying custom protein structures on phage surfaces, which can be enriched by affinity selection against the target.

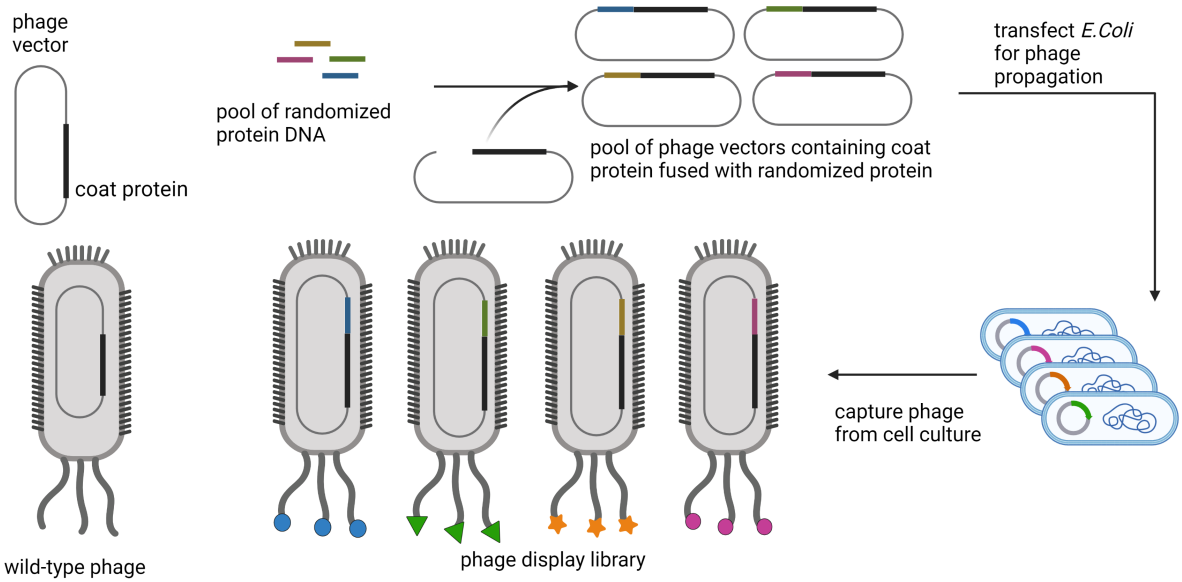

Figure 1. Phage display library construction. Pool of randomized protein DNA is fused with phage coat protein within a phage vector. The phage vector is transfected into the host and propagated in media. Phages are captured from the cell culture media. The constructed phage display library expresses the fused protein on the phage surface.

The phage display library is constructed by utilising recombinant DNA technology. Synthetic DNA fragments encoding either for antibody, protein or peptide are fused together with the phage coat protein within a phage vector (see Figure 1). Created phage vector is then delivered into *E.Coli* for propagation of the phage library (Lowman, 2013). Phage libraries are usually constructed by fusing sets of many randomized protein variants, effectively creating pool of phages where each phage displays only one random protein variant (Nakashima et al., 2000).

The technique has found applications in various research fields. For example, phage display has been used in vaccine development where antigen is expressed on the phage surface allowing to generate immune response in host for disease prevention or treatment (Chen et al., 2017). Other examples include phage display driven antibody discovery by enriching viral protein binding antibodies from randomized antibody-phage library pool using affinity selection (Oloketuyi et al., 2021; Barreto et al., 2019), antibody epitope mapping by displaying randomized peptides and enriching binders through affinity selection (Yang et al., 2022; Zhong et al., 2011; Palacios-Rodríguez et al., 2011). This work will focus on using peptide phage display for affinity selection against monoclonal antibodies (mAbs), which will serve as enrichment targets.

1.1.2 M13 Bacteriophage

Bacteriophages or phages are viruses that are capable of infecting and replicating only inside bacteria. They consist of viral coat surrounding the genetic material (DNA or RNA) that partially or fully encodes for its replication machinery and viral proteins. The viral coat contains structural proteins that protect the genetic material and functional proteins that participate in the host recognition and infection (White and Orlova, 2019).

Infection of the bacteria by phage is mediated by highly specific interactions between the host membrane proteins and the phage coat proteins. The range of hosts that can be infected by phage is determined by types of receptors located on the phage coat surface. Filamentous bacteriophages contain a tail composed of proteins that allows it to recognize and infect the host. Bacteriophages propagate by delivering viral genetic material into the host, where they can utilize the host's replication machinery and nutrients to self-propagate its genetic material and viral proteins. Synthesized structural proteins form a capsid and viral genetic material is packed inside. In general, phages can be classified based on their life cycle: lytic, lysogenic and non-lytic phages (Makky et al., 2021). Lytic phages, immediately after infecting the host start propagating its viral genetic material and expressing viral proteins. Newly synthesized molecules are assembled into phage virions, host is lysed and virions are released. Lysogenic phages enter dormant phase after host infection, during dormant phase they either integrate into

the host genome or remain as in plasmid form, lysogenic phase can last several hundreds of generations. During replication of the host viral genome gets carried over to the host progenies. When conditions are met lysogenic phages can enter in lytic cycle and aggresively propagate (Clokie et al., 2011). Non-lytic phages do not lead to cell death, after infection of the host they propagate, assemble and leak out of the host without heavily disrupting its membrane, hosts rate of growth is reduced (Moineau, 2013; Karimi et al., 2016).

M13 is a non-lytic rod shaped filamentous bacteriophage that infects gram-negative bacterial strains which display F-pili - flexible filaments on the cell surface. The M13 capsid is composed of around 2700 pVIII structural major coat proteins encircling the viral ssDNA. One end of the capsid contains five copies of pVII and pIX minor coat proteins, while the other end contains five copies of pIII and pVI minor coat proteins (Henry and Pratt, 1969; Simons et al., 1981) (see Figure 2). For the phage display applications all M13 structural proteins can be fused with a protein of interest (Sidhu et al., 2000; Fuh and Sidhu, 2000; Gao et al., 1999; Jespers et al., 1995). However, the pIII is the most commonly used one, since it can tolerate larger fusions and performs better in the affinity selection experiments (Iannolo et al., 1995; Loset et al., 2011). The pIII coat protein also mediates M13 infection, thus any fusion with pIII coat protein impacts phage capability to infect the host. This effect can introduce repertoire bias, especially when performing phage display assays with several rounds of affinity selection and phage propagation in bacteria (Juds et al., 2020; Matochko et al., 2012; Brammer et al., 2008). The synthetic phage genome often contains LacZα selection marker that can be used for blue-white colony screening.

The replication cycle of M13 bacteriophage begins with pIII coat protein of the phage attaching to the F-pili of *E. Coli*. Upon establishing contact with the viral pIII protein, the pilus retracts into the membrane where pIII then interacts with bacterial TolAQR complex. As a result of the interaction, the phage capsid destabilizes and pVIII coat proteins start dissociating into host's inner membrane, thus transferring packed phage ssDNA inside the host. Once inside the host, the phage genome is converted into the dsDNA by the machinery of the host (Olsen et al., 1972). New viral proteins and ssDNA copies of phage genome are synthesized, which in turn are utilized for further viral genome amplification. Synthesized viral pV protein acts as a negative feedback

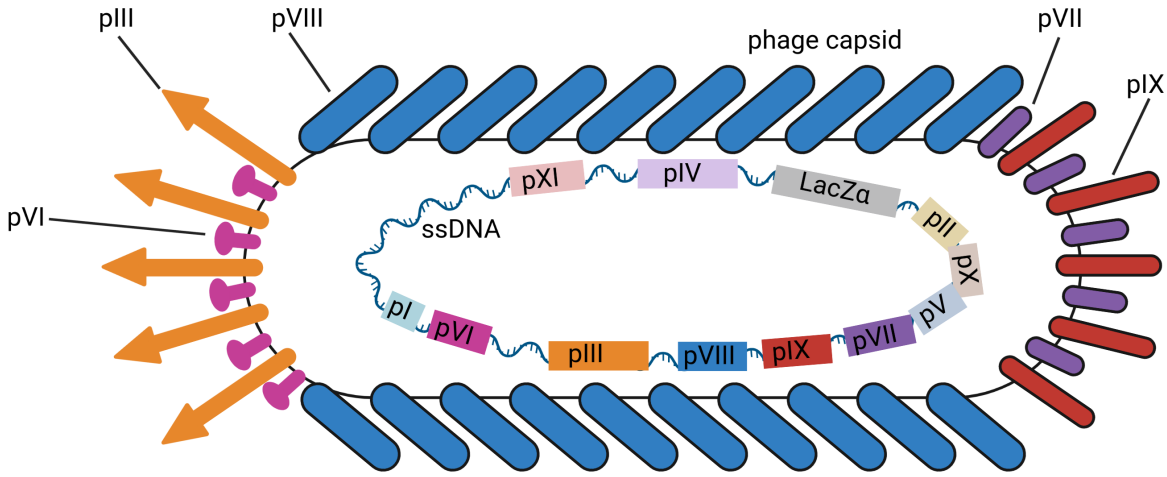

Figure 2. Schematic representation of components forming the M13 bacteriophage. The phage ssDNA is encapsulated by capsid formed from minor and major coat proteins.

signal and starts inhibiting viral ssDNA conversion to dsDNA by binding it. pV transports viral ssDNA to the host membrane, where assembly of the phage capsid takes place. Coat proteins pIX and pVII attach on one end of ssDNA, and as the ssDNA moves across membrane coat protein pVIII substitutes pV, fully encapsulating the viral DNA. The process terminates by attachment of pIII and pVI coat proteins, releasing phage particle (Bennett et al., 2011). The M13 replication cycle is illustrated in Figure 3.

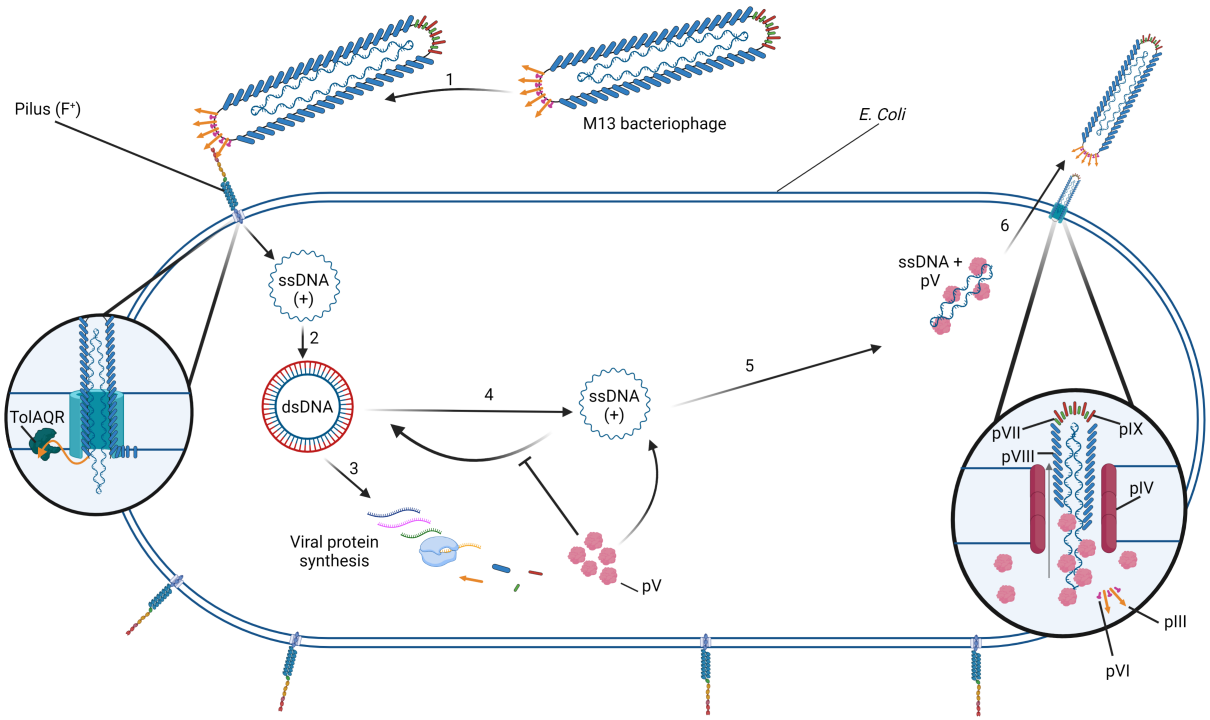

Figure 3. Schematic representation of M13 bacteriophage life cycle. 1) pIII phage coat protein attaches to F- pili displayed by the *E.Coli*. The pili retracts and pIII interacts with TolAQR membrane complex, destabilizing the phage capsid and releasing phage genome into the host. 2) Single stranded phage genome is converted into replicative form double stranded DNA. 3) Viral proteins get transcribed and synthesized by the host machinery. 4) From dsDNA, ssDNA phage genome gets synthesized, synthesized ssDNA can then again be turned into the dsDNA variant. 5) When pV protein is synthesized in high concentrations it inhibits dsDNA synthesis by packing around ssDNA and transporting it to the cell membrane. 6) As ssDNA moves through the pore formed by pIV protein, pVII and pIX proteins attach, pV gets displaced by pVIII attachment, at the end of ssDNA pIII and pVI proteins attach releasing phage from the cell.

1.1.3 Phage Display Biopanning

Biopanning is an affinity selection technique that selects biomolecules that exhibit affinity towards a given target. In context of the phage display, this technique aims to enrich phage or subset of phages displaying some biomolecule that binds most tightly to the target. Biopanning with phage display is often followed by sequencing step to determine the sequence of the displayed high-affinity biomolecule.

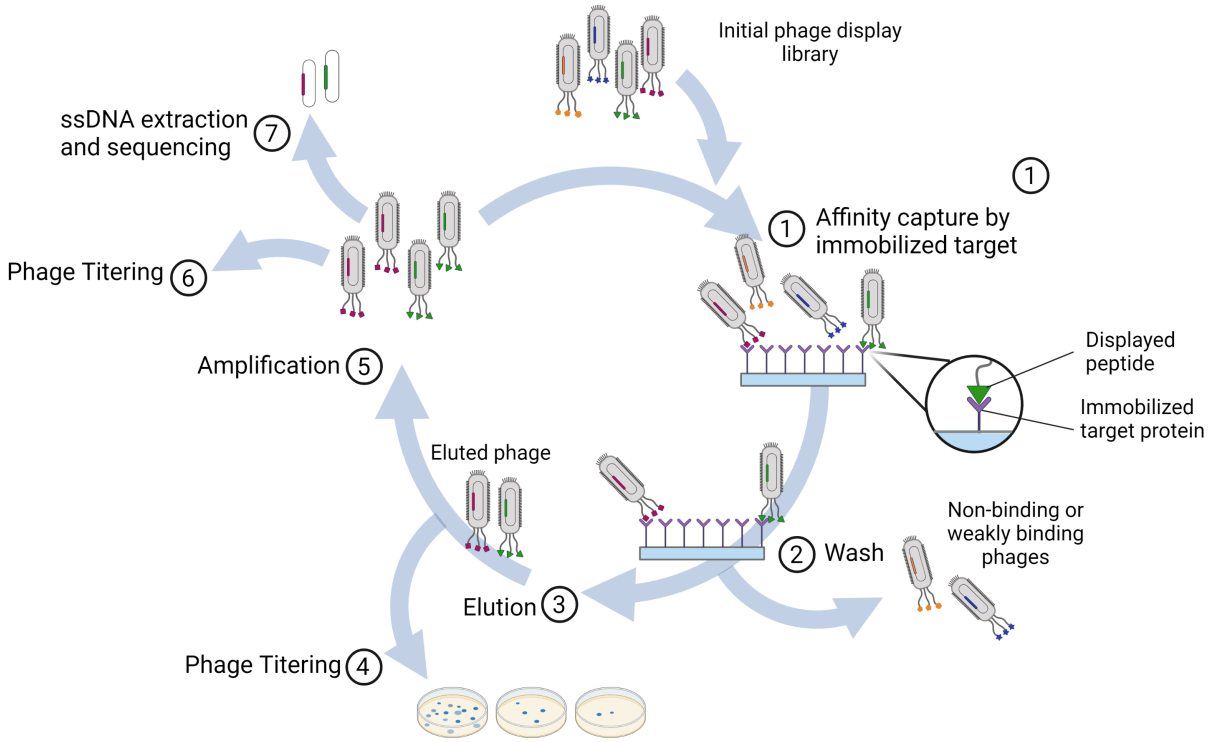

Figure 4. Schematic representation of phage display biopanning cycle. 1) Phages displaying protein with affinity toward the target bind. 2) Non-binding phages are washed. 3) Binding phages are eluted and 4) quantified by titering. 5) Phage eluate is amplified. 6) Pool of amplified phage is quantified by titering and 7) sequenced.

In the literature several *in vitro* phage display biopanning protocols exist. Most of the biopanning protocols can be divided into two classes based on phase at which the phage binds the target. Solution based phage-target binding protocols bind target with phage in solution, where target-phage complex is captured using beads that bind target affinity tag. Beads with bound target-phage complex are then recovered using magnets or centrifugation. In case of solid phase binding protocols, the target is first immobilized on the surface of polystyrene plate via non-specific interactions, or by biotinylating target and binding to streptavidin coated plate (McConnell et al., 1999; Koide et al., 2009). Immobilized target is incubated with solution of phage display library.

This project focused on solid phase phage-target binding technique.

Biopanning can be divided into four stages (see Figure 4): phage binding, washing, phage elution and phage amplification. During phage binding stage the immobilized target is incubated with phage display library, and the phages displaying peptides with affinity toward the target will bind the immobilized target. Binding stage is then followed by washing stage where weakly binding and non-binding phages are washed off by low concentrations of detergent, typically Tween–20. Remaining binding phage is then eluted using one or some combination of following methods: heavily adjusting pH of the environment, sonicating, adding reducing agents or proteases. Before amplification stage, phage elution is often quantified by phage titering. Eluted phage is then amplified using bacterial host, typically *E. Coli* (Willats, 2002). The amplified phage pool consists of phages displaying biomolecules that bind to target. This amplified pool of phages is then used as a starting phage library for the next panning round. Typically, 2-4 sequential rounds of biopanning are performed to enrich binding phages through directed evolution. After multiple selection rounds the amplified phage pools are sequenced to retrieve sequence of enriched protein or can be further studied by ELISA (Palacios-Rodríguez et al., 2011).

1.2 Protein sequence representations

Computational sequence representation methods aim to capture biologically relevant information from the protein sequence and represent it numerically. Early implementations of such techniques have focused on explicitly characterising the protein sequence by pre-defined residue properties (Guo et al., 2008; Shen et al., 2007). However, these methods are limited by the pre-defined features and do not capture the complexity of biological information encoded in the sequence. Therefore, deep learning methods that focus on sequence representation have emerged (Heinzinger et al., 2019; Alley et al., 2019; Asgari and Mofrad, 2015). Such methods train in a self-supervised manner on large protein datasets and learn complex features. The strength of self-supervised methods trained on sequence data is that they are not limited by the pre-defined constraints or lack of structural/labeled data.

Work of Heinzinger et al. (2019) has shown that deep learning based protein representations can carry information about protein family, localization and structure. Additionally, such representations can carry information relevant for predicting interaction between two proteins (Kimothi et al., 2019) or sequence mutation impact on function (Alley et al., 2019). There is no consensus in the literature regarding which deep learning representation method performs the best on certain task. In our work we apply several deep learning sequence representation algorithms to obtain representations from short-peptide sequences. And evaluate the ability of such methods to group specific target binding peptides in multi-dimensional latent space.

1.3 Dimensionality reduction

Dimensionality reduction algorithms aim to transform data from a high-dimensional space to a low-dimensional space while preserving as much of the relevant structure in the data as possible. Such algorithms are used for visualization of the high-dimensional data or for reducing number of features in order to save computing resources. Principal Component Analysis (PCA) algorithm is one of the most well-known of such algorithms (F.R.S., 1901). PCA works by geometrically projecting high-dimensional data onto lower dimensions called principal components (PCs). First PC is selected in a way that maximizes the variance across projected data points. Remaining PCs are selected in the same way with the additional constraint that they should be uncorrelated with previous PCs. However, PCA is effective only on data containing linearly correlated information and is sensitive to scale of units, thus requiring data scaling (Lever et al., 2017). Other dimensional reduction techniques, such as t-SNE and UMAP have addressed limitations of the PCA (van der Maaten and Hinton, 2008; McInnes et al., 2018).

t-SNE aims to preserve local and to some extent also global structure of the data. In brief, t-SNE starts with randomly scattering data points in low-dimensional space. It then starts an iterative process of reorienting data points, such that points that are close neighbours in the high-dimensional space would also be close in the low-dimensional space. One shortcoming of this method is that it does not capture global structure of the data accurately due to it being randomly initialized, i.e. it does not capture whenever two distant groups of points in low-dimensional space are also distant in high-dimensional space. t-SNE has been used with promising results in visualizing single cell transcriptomic data (Kim et al., 2020) and protein representations (Heinzinger et al., 2019).

1.4 Hierarchical clustering

Clustering algorithms are unsupervised algorithms that aim to discover patterns in the data and form clusters of data points in the feature space. Hierarchical clustering algorithm creates clusters with hierarchical property. Two types of such algorithms exist: agglomerative clustering and divisive clustering. This work will focus on using the agglomerative hierarchical clustering. The agglomerative hierarchical clustering works by first defining each data point to be in its own cluster. Then, by using some distance function, the algorithm finds for each cluster the nearest cluster and merges them together into one. It keeps merging the closest clusters until all clusters have been merged into one. As a result the algorithm has generated a hierarchical representation of data points, which often is shown in the form of a dendrogram (James et al., 2013). Usually the Euclidean distance is used to measure the distance between the points in the neighbouring clusters, while the linkage function defines the closest cluster using the distance metric. Several linkage functions exist. Single-linkage function finds the closest cluster by the shortest distance between the two closest points that reside in the two separate clusters. The complete-linkage function, however, finds the closest cluster by the shortest distance between the two furthest points that reside in the two separate clusters. Complete-linkage is the most commonly used, since it generally yields balanced results (James et al., 2013). In the field of biology, hierarchical clustering algorithms are often used to study and depict evolutionary relationships between species, proteins or viruses (Serrano-Solís and José, 2013).

1.5 k-Nearest Neighbours

k-Nearest Neighbours (KNN) is a supervised machine learning algorithm that uses the nearest set of existing labeled data to assign label to the unlabeled data, see Figure 5. The algorithm works in two stages - the first determines the closest k neighbours and the second determines label from the closest neighbours. During first stage it calculates the distance from unlabeled data point to all other labeled data points and picks k closest data points according to the distance function. During second stage it uses majority voting or weighted voting among the picked data points to assign label (Cunningham and Delany, 2020) based on neighbour labels. The most popular distance metrics used are Euclidean and Manhattan distances (Szabo, 2015).

Number of k determines how many neighbours will be used to determine the label. Using

very low k values can result in model over-fitting or on the contrary using too large k values in under-fitting. Over-fitting results from the model trusting labeled data completely, thus noise existing in the data would negatively impact predictions. On the other hand, under-fitting results from model being too invariant towards labeled data, resulting in poor boundary definition as depicted in Figure 5.

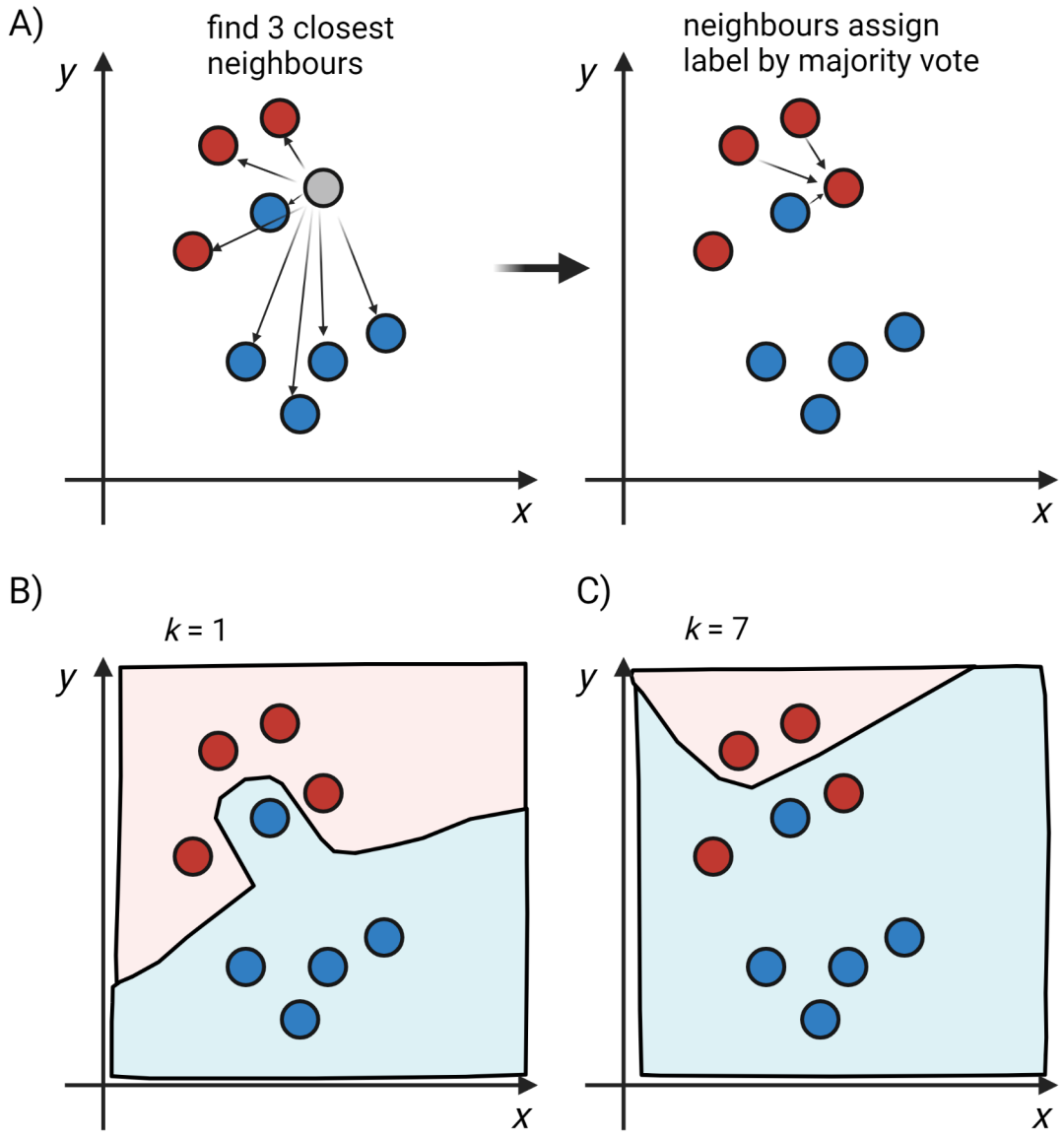

Figure 5. Visualization of KNN algorithm. A) Distances to all labeled data points (labeled by color) from unlabeled data point are calculated. Closest k neighbours (in this case 3) are selected and by majority vote they assign a label. B) Shows the case of over-fitting with $k = 1$. Cluster of the red points contains blue boundary zone due to noisy blue data point. The background color boundary depicts the label that would be assigned to unlabeled data point. C) Shows the case of under-fitting with $k = 7$. Majority of the red point cluster resides in the blue boundary zone because algorithm is not sensitive enough to create red zone boundary encircling red point cluster due to large k value.

1.6 Biopanning Data Bank

Biopanning Data Bank (BDB) is an expert curated database that aims to store peptide phage display experimental results from published articles. As of the latest release (December 30, 2020), BDB has accumulated 3562 biopanning datasets from 1697 published articles. In total it contains 33097 peptide sequences stored from 2156 targets. Most of the sequences in biopanning datasets are obtained by Sanger sequencing resulting on average in 20-100 total peptides listed per target. Considering that the majority of phage display libraries are with approximate complexity of 10^9, and eluate complexities from affinity selection can still be in ranges of hundreds of thousands, Sanger sequencing method covers only a fraction of binding peptide sequence diversity space (He et al., 2015, 2018). The databank provides following information for the most of the stored datasets: starting complexity of phage library used, number of affinity selection round the dataset is obtained from, description of phage library used, sequencing method used, reference to research article. Example of the dataset entry in the BDB is given in Table 1.

Table 1. Example of few database entries from BDB collection.

Target name	Peptide sequences	Number of unique sequences	Panning round	Library name	Library complexity	Sequencing method	Article reference
Anti-Ogawa O-antigen monoclonal antibody S-20-4	SHKLHVK SHRLPLK SHRLPAK SHRLPVK	4	3/4	Ph.D.-7	10^9	Sanger	17881351
Gitoxin	RDFYYN GERFFN SKRYIN	3	6/6	X6 phage display	10^7	Sanger	7499368

2 THE AIMS OF THE THESIS

Several machine learning methods exist in the literature for extracting protein sequence representations; however, no data is available about their performance in meaningfully representing short peptide sequences. Therefore, this work aims to investigate the capabilities of the current deep-learning-based protein representation techniques in the context of mapping epitope to the paratope.

Current experimental methodologies that aim to retrieve monoclonal antibody-binding short protein sequences are limited by their design, capturing only a fraction of binding sequences. Therefore, the second aim of this work is to establish a standardised experimental procedure for acquiring reliable large-scale data that captures the sequence diversity recognized by the paratope.

3 EXPERIMENTAL PART

3.1 MATERIALS AND METHODS

3.1.1 Bacterial strain maintenance

Unthawed glycerol stock of supplied *E.Coli* ER2738 bacterial strain was scraped with sterile pipette tip and streaked onto a LB+Tet (20 µg/ml) agar plates. The plate was wrapped in parafilm and incubated at 37°C for 14 hours. Plate was stored at 4°C in dark and used as a bacterial colony source. Single colony was picked from the plate and inoculated in 250 mL sterile Erlenmeyer flask containing 20 mL of LB+Tet (20 µg/ml) media. Cell culture was incubated at 37°C in rotary shaker set at 250 rpm until OD_{600} reached approximately 0.5. Bacterial glycerol stocks were prepared by dispensing 500 µl of bacterial culture into cryotubes containing 500 µl of 1:1 glycerol MilliQ water solution. Cryotubes were frozen in a freezing container and stored at -80°C.

3.1.2 Phage amplification

To amplify Naive phage library, 100 mL of LB+Tet (20 µg/ml) media was inoculated with ER2738 colony from LB+Tet (20 µg/ml) plate in a 250 mL Erlenmeyer flask and incubated in rotary shaker set at 37°C and 250 rpm until OD_{600} reached approximately 0.05. 1 µl of Naive phage library was diluted 1000 times, from diluton 1 µl containing 10^7 phage particles was used to infect bacterial culture at MOI of approximately 0.1. The bacterial culture was incubated for 5 hours in rotary shaker set at 250 rpm and 37°C.

To amplify the phage eluate from biopanning rounds, 20 mL of LB+Tet (20 µg/ml) media was inoculated with ER2738 colony from LB+Tet (20 µg/ml) plate in a 250 mL Erlenmeyer flask and incubated in rotary shaker set at 37°C and 250 rpm until OD_{600} reached approximately 0.01 to 0.05. When culture reached the desired OD, eluted phage was added and incubated for 5 hours in rotary shaker set at 37°C and 250 rpm.

To purify amplified phage, grown cell culture with phage was transferred to 50 mL Falcon tube and centrifuged for 30 minutes at 5000 g and 4°C, supernatant was transferred to a fresh

tube and re-centrifuged. The 80% of upper supernatant volume was then transfered to a fresh Falcon tube to which 20% PEG/2.5 M NaCl solution was added in amount of $1/6^{th}$ of the supernatant volume. Phages were precipitated overnight at 4°C. Tube with phage precipitate was centrifuged at 5000 g for 30 minutes at 4°C, supernatant was discarded. Precipitated phage pellet was resuspended in 1 mL of TBS(50 mM Tris–HCl, 150 mM, pH 7.5). Supernatant was transfered to a microcentrifuge and centrifuged at 14'000 rpm for 5 minutes at 4°C to pellet residual cells. Supernatant was transferred to a fresh microcentrifuge tube and phage was re-precipitated by adding 170 µl of 20% PEG/2.5 M NaCl and incubating on ice for 2 hours. Tube with phage precipitate was microcentrifuged at 14'000 rpm for 10 minutes at 4°C, supernatant was discarded, pellet was resuspended in 200 µl of TBS and transferred to fresh microcentrifuge tube. Finally, tube was centrifuged at 14'000 rpm for 1 minute, supernatant was transferred to a fresh LoBind microcentrifuge tube and stored at 4°C.

3.1.3 Phage titering

For determining approximate number of phages in a sample, 10 mL of LB+Tet (20 µg/ml) media was inoculated with ER2738 colony from LB+Tet (20 µg/ml) plate in a 250 mL Erlenmeyer flask and incubated in rotary shaker at 37°C and 250 rpm until OD_{600} reached approximately 0.5. Serial dilutions of phage sample with 100 µl of total volume were prepared in following ranges with 10-fold increments: 10 to 10^6-fold dilutions for eluted phage samples, 10^8 to 10^{12}-fold dilutions for amplified phage samples. 200 µl of bacterial culture grown to desired OD_{600} was dispensed into microcentrifuge tubes (per each dilution), and 10 µl of corresponding phage dilution was added, vortexed, and incubated at room temperature for 5 minutes allowing phage to infect bacteria. Contents of microcentrifuge tube were then pipetted to culture tubes containing 3 mL of warm (55°C) Top agar, vortexed thoroughly, and immediately poured onto pre-warmed LB/IPTG/X-gal agar plates, plate was tilted and rotated to spread the Top agar evenly. Plates were cooled for 5 minutes in a laminar flow cabinet, then incubated inverted overnight at 37°C. Plates containing approximately 100 plaques were used to count number of blue plaques. Counted plaques were multiplied by total dilution factor to derive phage titer in plaque forming units (pfu) per 1 µl.

3.1.4 Surface panning procedure

To enrich peptides binding to selection of monoclonal antibodies, Ph.D.–12 Phage Display peptide library (NEB #E8110S, lot #10111203) was used. To increase stringency of peptide selection, three consecutive panning rounds were performed with increased Tween–20 concentration in wash buffer and decreased monoclonal antibody concentration added in coating solution for each subsequent panning round. 150 µl of 0.1 M $NaHCO_3$ coating solution containing 100 µg/ml of antibody was added to the well of the 96–well Nunc MaxiSorp plate (ThermoFisher #442404). Three wells were prepared containing following monoclonal antibodies: Anti–p53 antibody PAB 240 (Abcam #ab26), Anti–SARS–CoV–2 Spike Glycoprotein S1 antibody CR3022 (Abcam #ab273073) and Anti–Influenza A H1N1 hemagglutinin antibody C102 (Abcam #ab128412). Additionally, two control wells were filled with 150 µl of 0.1M $NaHCO_3$. Plate was left sealed at 4°C overnight on see-saw rocker set at 7 tilts a minute. Next day, plate was firmly slapped onto a paper towel to remove coating solution. Each well was then filled with 250 µl of blocking solution (0.1 M $NaHCO_3$ + 5 mg/mL BSA) and incubated for 2 hours at 4°C. After incubation, blocking solution was removed by firmly slapping the plate onto a paper towel and washed 6 times with TBST (TBS + 0.1% Tween–20). 100 µl of TBST containing 10^{11} of phage particles were pipetted onto each well coated with antibody and on one control well coated with BSA, the second control well was instead filled only with TBST and served as control for phage cross-contamination. Plate was left at room temperature on rocker set at 250 rpm for 1 hour, then washed for 10 times with TBST removing unbound phages. Bound phages were eluted by adding to each well 100 µl of elution buffer (0.2 M Glycine–HCl, pH 2.2), gently rocking for 5 minutes, then collected in separate LoBind tubes containing 15 µl of neutralizing buffer (1 M Tris-HCl, pH 9.1). 1 µl of sample from each well was used for determining eluate phage titer. Eluates from wells where phage libraries were added were further amplified in bacteria.
Subsequent two panning rounds were performed using respective amplified phage libraries from previous round and panned against the same target. Tween–20 concentration in TBST was raised to 0.3% and 0.5% while antibody concentration in coating solution was decreased to 66 µg/mL and 33 µg/mL respectively.

3.1.5 NGS library preparation

Following phage samples were selected for the sequencing: Naive phage library, Naive amplified phage library, amplified antibody binding phages from each panning round, and amplified phages panned against BSA coated control after each panning round. 10 µl of each phage sample was diluted with 50 µl of lysis buffer from the kit, and phage ssDNA was extracted using QIAprep Spin Miniprep Kit (Qiagen, #27104) following the QIAprep M13 kit protocol. Eluted ssDNA concentration was determined using ThermoFisher 2000 NanoDrop spectrophotometer. ssDNA region encoding for phage displayed peptide was amplified by 3-step 25 cycle PCR using NEBNext® Ultra™ II Q5® high-fidelity polymerase Master Mix (NEB #M0544S), Table 2 summarizes PCR run conditions. Forward and reverse primers used in the PCR were designed using amplicon phasing method (Wu et al., 2015) by including non-complementary spacers with length of 1 to 3 bases and the Illumina adapter overhang sequences. Added spacers shifted amplicon sequences resulting in higher base read diversity in each sequencing cycle. Base diversity allowed sequencing machine to better distinguish separate sequence clusters on the sequencing chip, therefore, allowing to achieve higher quality in sequencing run. Figure 6 summarizes the PCR primer design for low-diversity amplicon libraries. Table 3 summarizes primer sequences used for each sample. 1 µl of the PCR product was used for 2% agarose gel electrophoresis, gel was post-stained with GelRed® (Biotium, #41003-T) and band sizes were confirmed by UV imaging.

Table 2. PCR run conditions for generating target region amplicons with Illumina adapters.

Step	Temp.	Time	
Initial denaturation	98°C	60s	
Denaturation	98°C	10s	
Annealing	64°C	30s	25x
Extension	72°C	20s	
Final extension	72°C	60s	

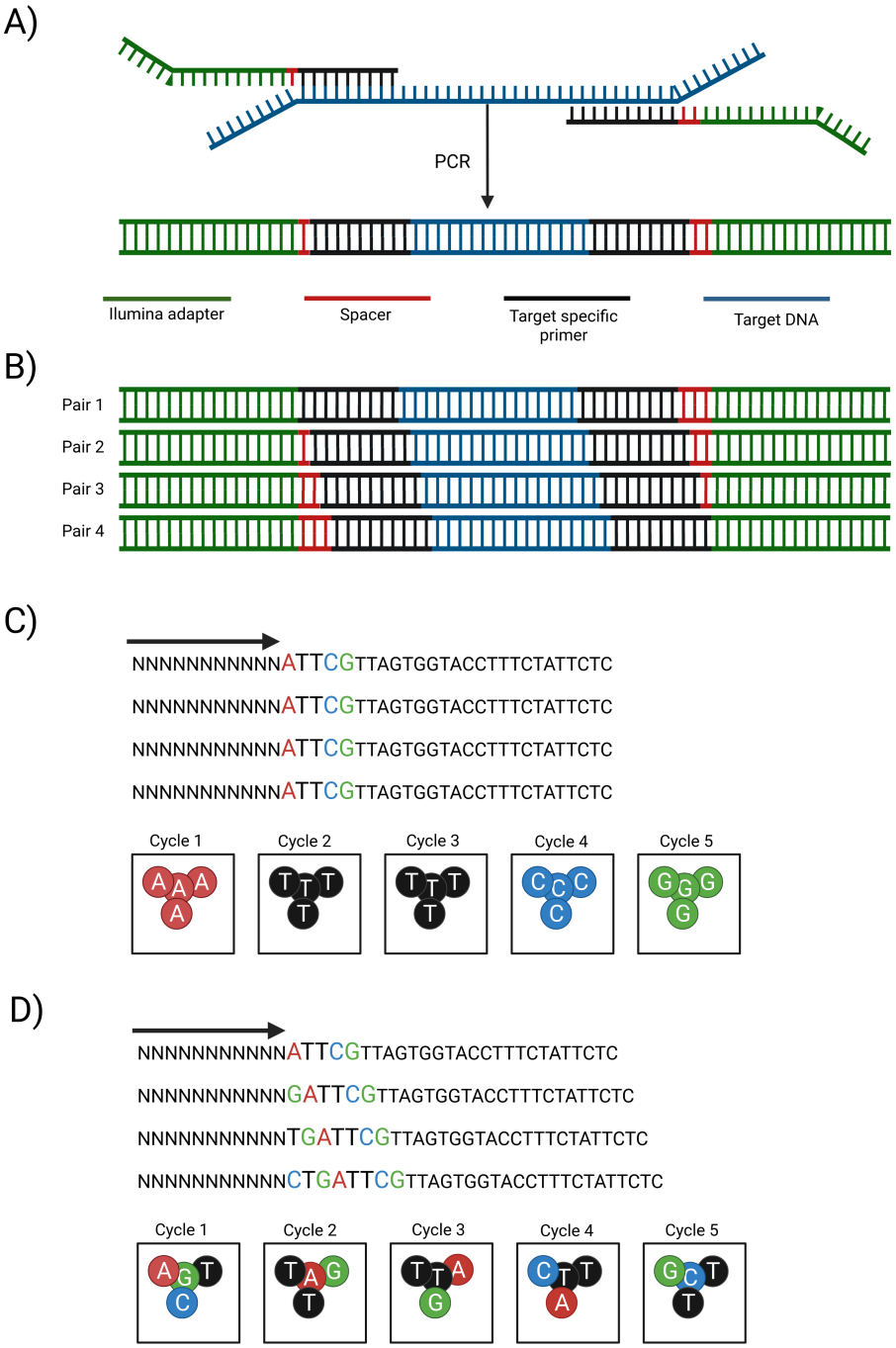

Figure 6. Summary of targeted PCR design. A) Targeted PCR using phage ssDNA as a template, Illumina specific overhangs and spacer nucleotides are added. B) Product of the PCR reaction with different primer pairs, resulting in forward and reverse target sequence shift due to added spacers. C) Example of first 5 Illumina sequencing cycles on non-phased low-diversity sample. Clusters on a chip are hard to seperate. Bases are color-coded. D) Example of first 5 Illumina sequencing cycles on phased low-diversity sample. Clusters on a chip are easier to seperate, resulting in a higher read quality and in a higher amount of total reads from sequencing run. Bases are color-coded.

Table 3. PCR primer pair distribution across samples.

Sample name	Forward primer	Reverse primer
Naive non-amplified	Fw1-pair1st	Rw1-pair1st
Naive amplified	Fw1-pair2nd	Rw1-pair2nd
CR3022 Panning 1	Fw1-pair3rd	Rw1-pair3rd
H1N1 Panning 1	Fw1-pair3rd	Rw1-pair3rd
PAB240 Panning 1	Fw1-pair3rd	Rw1-pair3rd
Control (BSA) Panning 1	Fw1-pair3rd	Rw1-pair3rd
CR3022 Panning 2	Fw1-pair4th	Rw1-pair4th
H1N1 Panning 2	Fw1-pair4th	Rw1-pair4th
PAB240 Panning 2	Fw1-pair4th	Rw1-pair4th
Control (BSA) Panning 2	Fw1-pair4th	Rw1-pair4th
CR3022 Panning 3	Fw1-pair4th	Rw1-pair4th
H1N1 Panning 3	Fw1-pair4th	Rw1-pair4th
PAB240 Panning 3	Fw1-pair4th	Rw1-pair4th
Control (BSA) Panning 3	Fw1-pair4th	Rw1-pair4th

3.1.6 NGS

Pre-processing of the PCR sample and sequencing was carried out by DNA Sequencing facility in the Department of Biochemistry, University of Cambridge, on Illumina MiSeq platform using 75 bp paired-end sequencing. Sample library preparation and sequencing was done according to Illumina 16s metagenomic library preparation guide (Illumina, San Diego, USA). Briefly, the PCR product was first cleaned up using AMPure XP beads in 1:1 bead sample ratio. Then indices for multiplexing and Illumina sequencing adapters were attached by Index PCR using Nextera XT Index kit (Illumina, San Diego, USA). The PCR product was cleaned up using AMPure XP beads in 1:1 bead sample ratio and validated using Agilent 2100 Bioanalyzer instrument. Then normalized by diluting samples to equimolar concentrations. The samples were pooled together, denatured and loaded onto a sequencing chip. 35% of PhiX spike-in was added, and total cluster density was set at 948 K/mm^2 to increase sequencing quality of the low-diversity amplicon libraries. The PhiX reads were removed by Illumina software, remaining reads were de-multiplexed into separate .fastq files per sample. Forward and reverse reads were stitched prioritising higher quality base read in MiSeq Reporter software v2.6.2.3.

3.1.7 Sequencing data processing

For extracting displayed peptide sequence raw reads were pre-processed in Python. FASTQ files containing DNA read sequences were first filtered using regex pattern searching method

(Aho, 1991). Regex expression searched for two 6bp constant regions that are flanking 36bp DNA sequence encoding for phage displayed peptide. Upon expression match DNA read was truncated to only contain DNA regions encoding for the displayed peptide. Fragments were filtered to be 36bp long and to not contain any stop codons. According to the phage library supplier codon pattern in the fragment should be NNK (N - A, C, T, G; K - G or T), fragments without NNK pattern were discarded. Fragments were quality filtered removing fragments with any bases having phred score below 30. DNA fragments were then translated to the 12-mer peptide sequence using Biopython (Cock et al., 2009) package with NCBI codon table 11. Protein sequences in mAb target samples were compared with sequences in all three control samples panned against BSA, and matching sequences were discarded from mAb target sample. Additionally, peptides that were present across samples panned against different mAb targets were also removed to further reduce number of non-specific peptide sequences.

3.1.8 Sequencing data normalization

Sample normalization was carried out using Naive amplified NGS sequencing data sample as a reference. It was done according to the Equation (1).

$$n_{sample\,normalized} = \frac{n_{sample} \cdot ref_{total}}{n_{ref} \cdot sample_{total}} \tag{1}$$

Where n_{sample} is the copy number of a sequence in a sample, n_{ref} is the copy number of the same sequence in the reference (Naive amplified). ref_{total} and $sample_{total}$ is the total number of sequences (including clones) in reference and sample respectively. $n_{sample\,normalized}$ is the resulting sequence expressed in normalized abundance quantities. Normalization step requires that sequence occurs at least once in both the reference and the sample sequencing results. To avoid loss of large quantities of possibly relevant sample sequences due to limited sequencing depth of the high diversity Naive amplified library, pseudo value of 1 was used for n_{ref} if exact reference sequence could not be matched with sample sequence.

3.1.9 Hierarchical clustering

Global alignment of protein sequences was carried out using Needleman-Wunsch algorithm without penalizing terminal gaps, allowing the alignment of different length sequences. The pairwise alignment score was calculated for each peptide pair using Scikit-Bio (Scikit-Bio, 2020) Python package. Hierarchical clustering was done using complete linkage method together with Euclidean distance using SciPy Python package (Jones et al., 2001).

3.1.10 Motif discovery

To identify motifs from sequencing data MEME motif discovery algorithm was used (Bailey et al., 2009). Algorithm was applied to .fasta files containing peptide sequences. Following run conditions were used: motif occurrence - zero or one motif per sequence, minimum width of motif - 3 residues, minimum number of sequences containing the motif - 7. Top 100 normalized sequences from samples panned against mAbs and BSA were used to extract motif logos and occurrence counts for each of the motif found.

3.1.11 Generating sequence representations

Pre-trained self-supervised deep-learning models were used to generate protein sequence representations. Various self-supervised models from the field of natural language processing (NLP) and the field of protein structure prediction were chosen based on the availability of open-sourced model implementations. Protein sequence was fed into the model and generated representation was stored in a dataframe using custom Python script.

NLP skip-gram based BioVec model that has been pre-trained on Swiss-Prot protein database was used to generate 100-dimensional protein sequence representations (Bairoch and Apweiler, 2000; Asgari and Mofrad, 2015). The model was implemented using code from public repository (Kyu Ko, 2016). An alternative implementation of this model was also used (Changgeon Lee, 2017). Within the framework of this project this model variant is referred to as ProtVec.

Masked language modelling (MLM) based SeqVec model that has been pre-trained on UniRef50 protein database was used to generate 1024-dimensional representations (Suzek et al., 2007; Heinzinger et al., 2019). Publicly available implementation was used (Dallago et al., 2021).

Another MLM based UniRep model trained on UniRef50 database was used to generate 1900-

dimensional representations (Suzek et al., 2007; Alley et al., 2019). UniRep pre-trains by iterating over the protein sequence and predicting the next residue. To see if any performance gains can be achieved by different representation generation possibilities, three types of representations were generated. h_ avg that averages all hidden state values from the model as it infers each sequence residue, h_ final that takes hidden state value after whole sequence has been parsed and c_ final that takes final cell state value. Implementation from public repository was used (Surge Biswas, 2019).

Structure prediction based RGN model trained on ProteinNet dataset was used to generate 800-dimensional representations (AlQuraishi, 2019b,a). Model pre-trains by sequentially encoding each residue bidirectionally and then generating a protein 3D structure that is compared to an existing structure. Four possible representations from model were extracted per each input sequence. RGN_ f1, RGN_ f2 that are representations from two forward hidden layers and RGN_ r1, RGN_ r2 that are representations from two reverse layers. Implementation from publicly available code repository was used (AlQuraishi, 2018).

BERT is an MLM based model pre-trained on merged UniRef and BFD protein datasets (Suzek et al., 2007; Jumper et al., 2021; Elnaggar et al., 2020). 1024-dimensional representations were extracted by using public implementation of the model (Dallago et al., 2021).

PLUS is another MLM based model that has been pre-trained on Pfam dataset (Finn et al., 2015; Min et al., 2019). 1024-dimensional representation were extracted by using public implementation of the model (Dallago et al., 2021).

3.1.12 k-Nearest Neighbours

KNN classification method was used to generate Naive and alignment based performance baselines. It was also used to evaluate relative performance of representation methods in comparison with aforementioned baselines. Leave one out cross-validation technique was used to record for each sequence the predicted and the actual target value (James et al., 2013). Then the weighted average accuracy score was calculated by taking the average of the fraction of the correctly predicted target values in each target value class. Algorithm was re-run with different number of neighbours (k) to test its robustness. KNN was implemented and the results were generated using Python Scikit-Learn machine learning package (Pedregosa et al., 2011).

KNN alignment baseline was generated using Needleman-Wunsch sequence alignment algorithm as a scoring function to determine the closest neighbours. To make prediction to which target the sequence belonged to, the sequence was first aligned to every other sequence in the dataset then alignment scores were ranked from the highest to the lowest. Top k scores based on the number of neighbours were selected and prediction was made based on which targets the neighbours belonged to. In case majority of the neighbours were from one target, the algorithm predicted the target shared across majority of neighbours. In case no majority of neighbours were from one target the algorithm choose at random the neighbour and used its assigned target for the prediction.

Naive baseline was based on using neighbours that had random target associated with them. For each prediction k random neighbours were generated, each neighbour had random target value associated with them. The randomising function that assigned random target values followed the same distribution as the distribution of the target values in the dataset, i.e. in case one target was more prevalent in the dataset the randomising function assigned larger probability towards selecting it. The prediction was made based on neighbour target values, similarly as in KNN alignment baseline.

KNN applied on full and dimensionality reduced representations used Euclidean distance as a scoring function to determine k closest neighbours. Top k neighbour targets were used to determine the prediction in the same manner as mentioned previously.

3.2 RESULTS

3.2.1 Dataset A

As a first step in this project a dataset was constructed, containing peptide sequences that bind specific monoclonal antibody targets. Monoclonal antibodies were chosen due to their specific target protein binding capabilities, thus any peptides bound in the following experiments would mimic the epitope that can be recognized by the antibody. The BDB database was used as the source for antibody-binding protein sequences, as it contains curated publicly available experimental data from affinity selection assays. Dataset A was constructed by picking top 6 mAb targets according to number of unique peptide sequences binding them. In total, 418 unique peptides were merged from 19 biopanning experiments. Summary of Dataset A is given in Table 4. Peptide sequences from cyclic libraries were pre-processed removing flanking cysteine residues. This was done to avoid any bias introduced in alignment scoring, where flanking cysteine residues would be scored higher than the actual target binding peptide motif. Summary of multi-sequence alignment (MSA) logos for each of the mimosets is given in the Appendix Figure 18, detailed mimoset phage library description is given in Appendix Table 18, full target names are given in Appendix Table 19.

Before evaluating protein representation techniques on the dataset, it was important to establish a simple and robust performance baseline that would score similarity between peptides and group related peptides together. One way to establish such a baseline is to rely directly on the peptide sequence. Therefore, pairwise alignment of peptide sequences was carried out across the dataset. To group similar sequences together, hierarchical clustering was applied to alignment scores. For developing this baseline, an assumption was made that the peptide set which is binding a target would contain a strong sequence motif shared between binders. Thus this baseline was expected to result in clusters that contain binding peptides only from one target. From resulting dendrogram in Figure 7, cutoff value of six clusters was chosen that minimized the number of clusters while maximising the number of sequences in cluster belonging to one target. Distribution of peptides across clusters was analyzed in Table 5, a detailed table containing each mimoset allocation to cluster can be viewed in Appendix 20. In clusters 2 and 3 the results follow the baseline assumption, respective peptides from target 3 and

Table 4. Description of mimosets in the Dataset A. Motifs are based on MSA logos from Appendix Figure 18

Target	Mimoset	Number of unique peptides, n	Panning Cycle	Library type	Peptide length	Motif
1	1.1	2	4	Linear	14	[W/F]SDL
1	1.2	32	3	Linear	15 or 21	-
1	1.3	19	-	Circular	12	WSD
1	1.4	3	3	Linear	12	-
1	1.5	2	3	Linear	22	-
1	1.6	8	-	Circular	10	SDL
1	1.7	10	-	Circular	14	WSDL
2	2.1	18	3	Linear	15	-
2	2.2	41	3	Circular	13	-
3	3.1	90	3	Linear	9	[W/F]RxRLL
4	4.1	9	3	Linear	16	G[W/F]A
4	4.2	6	-	Linear	20	G[W/F]A
4	4.3	20	1	Linear	7	-
4	4.4	17	2	Linear	7	G[W/F]A
4	4.5	8	3	Linear	7	GxA
5	5.1	44	3	Linear	7	DKW
5	5.2	27	3	Circular	12	DKWA
6	6.1	28	4	Linear	10	SDLxKL
6	6.2	34	5	Linear	10	SDLxKL

6 were clustered into separate clusters while also retaining at least 85% of their specific binding peptides. Although cluster 1 contained peptides only from target 5, majority of target 5 peptides were present in cluster 5 intermixing with target 4 binders. It was expected that peptides from mimoset 1.1, 1.6 and 1.7 would be grouped together with peptides from mimoset 6.1 and 6.2 due to shared SDL motif. However, upon examining cluster 3, only six of target 1 peptides with SDL motif were present, while the remaining 17 peptides from target 1 with SDL motif were grouped in cluster 6 together with only two peptides with SDL motif from target 6.

Table 5. Target peptide distribution (in fractions) across clusters in the Dataset A.

Cluster	1	2	3	4	5	6	Total peptides, n
Target							
1			0.08	0.53		0.39	76
2				0.90	0.05	0.05	59
3		0.96				0.04	90
4				0.65	0.23	0.12	60
5	0.38				0.59	0.03	71
6			0.85	0.10		0.05	62

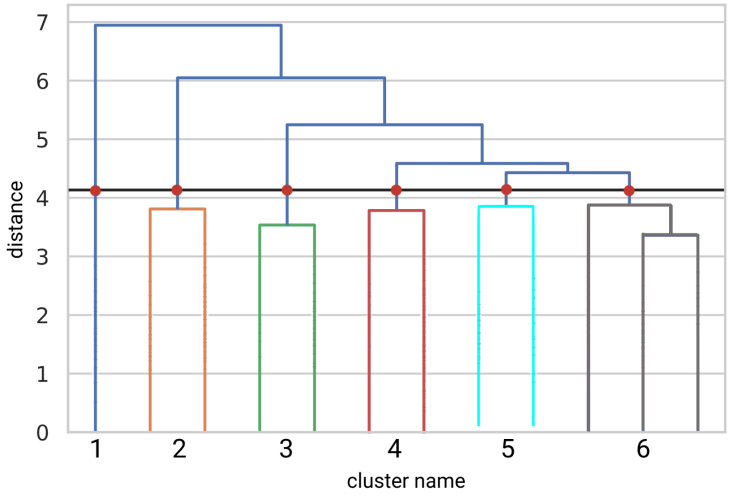

Figure 7. Hierarchical clustering dendrogram. Hierarchical clustering of peptides in Dataset A based on sequence pairwise alignment score. The resulting dendrogram is split into six clusters with chosen threshold. Vertical axis represents the distance between clusters, horizontal axis specifies cluster name linked to the color code.

These results demonstrate that even with shared motifs peptides could be fractionated across different clusters. Target peptide intermixing in one cluster and the spread of target specific peptides across multiple clusters shows that hierarchical clustering based on sequence alignment can at best map only lesser fraction of peptides to their binding targets in single cluster. We did

not find clustering threshold that could cluster together single target peptides without increasing number of clusters significantly. Hierarchical clustering on pairwise alignment resulted in a weak and non-robust baseline for grouping peptides binding the same target. Therefore, another baseline using KNN algorithm was constructed. In doing so, the assumption that specific target binding peptides should strictly share a degree of similarity between themselves was abandoned. Instead a weaker assumption was made that within a set of binding peptides different sub-groups of similar sequences exist, thus KNN metric would be more suitable at assigning these to a specific target.

KNN baseline was constructed by substituting traditionally used Euclidean distance for calculating nearest neighbours with Needleman-Wunsch alignment score. This method allowed to establish a prediction accuracy baseline that uses sequence alignment score. Performance of the KNN algorithm was evaluated using leave one out cross-validation method and the method robustness was tested using different number of neighbours. Weighted average accuracy for the sequence similarity baseline ranged from 83% to 84% (see Table 6).

Table 6. Weighted average accuracy values for KNN algorithm with different number of neighbours when predicting on Dataset A. KNN distance function was based on sequence alignment scores. Accuracy values were calculated using leave one out cross-validation.

	KNN number of neighbours, k				
	1	3	5	10	15
Alignment baseline	0.83	0.83	0.84	0.84	0.84

The promise of pre-trained protein sequence representation methods is their ability to generate representations that capture biological properties of a protein from its sequence in a high-dimensional vector. Such representations are difficult to interpret and to visualize due to their high-dimensionality. To visualize the behaviour of computational representations and evaluate their ability to group target specific peptides, t-SNE dimensionality reduction algorithm was applied to generated representations. Peptide sequence representations were generated using various publicly available pre-trained self-supervised models from the literature (Ibtehaz and Kihara, 2021). Computed peptide representations were reduced to two dimensions for visualization. Some of the visualizations are depicted in Figure 8, visualization of remaining representation methods can be seen in Appendix Figure 19.

To quantify which representation method allows for better classification performance, KNN

classification algorithm was applied to dimension reduced (two dimensions) and full dimension protein sequence representations, and the distance between neighbours was calculated using Euclidean distance. Additionally, a Naive KNN baseline was established by measuring the KNN performance on the randomized data. Table 7 summarizes KNN performance on different representations. From the results it could be seen that some fully dimensional representations on average performed marginally better than reduced representations with some exceptions. All representation methods resulted in better performance than the Naive baseline. It is interesting to observe that large deep learning based models like SeqVec, BERT, PLUS, RGN, UniRep generally performed worse or on par with comparably small and light BioVec representation model. Performance values across different number of neighbours changed only slightly showing robustness of the method. UniRep c_final and BioVec representations led to a marginally better KNN performance than the alignment baseline.

Table 7. Weighted average accuracy values for KNN algorithm with different number of neighbours when predicting on different types of representations. Accuracy values were calculated using leave one out cross-validation. Representations were generated on Dataset A.

	Full dimensionality					Dimensionality reduced				
	KNN number of neighbours, k					KNN number of neighbours, k				
Representation	1	3	5	10	15	1	3	5	10	15
ProtVec	0.80	0.80	0.77	0.79	0.78	0.79	0.79	0.76	0.74	0.71
BioVec	0.85	0.82	0.83	0.82	**0.83**	0.84	**0.85**	**0.85**	**0.84**	**0.85**
UniRep c_final	**0.87**	**0.85**	**0.86**	**0.84**	**0.83**	**0.88**	0.84	0.84	**0.84**	0.81
UniRep h_final	0.68	0.63	0.65	0.64	0.66	0.67	0.66	0.62	0.60	0.56
UniRep h_avg	0.74	0.74	0.76	0.73	0.73	0.74	0.73	0.72	0.69	0.67
SeqVec	0.82	0.81	0.82	0.81	0.80	0.85	0.83	0.84	0.82	0.81
RGN_f1	0.80	0.82	0.82	0.80	0.79	0.83	0.82	0.81	0.78	0.74
RGN_f2	0.82	0.83	0.82	0.79	0.77	0.81	0.84	0.82	0.80	0.78
RGN_r1	0.70	0.73	0.75	0.71	0.73	0.76	0.72	0.70	0.68	0.67
RGN_r2	0.70	0.70	0.68	0.68	0.65	0.69	0.66	0.62	0.62	0.57
BERT	0.72	0.68	0.68	0.69	0.70	0.73	0.67	0.69	0.64	0.64
PLUS	0.59	0.61	0.64	0.61	0.59	0.61	0.58	0.56	0.52	0.52
Naive baseline	0.20	0.17	0.14	0.15	0.17	0.20	0.17	0.14	0.15	0.17

In order to investigate potential sources of bias and information leakage in the dataset, plots of dimensionality reduced representations were assesed. Plotted BioVec dimension reduced representations(see Figure 8) were studied in detail by examining peptide sequences in the groups. It was observed that the cluster which consists of a mix of peptides from targets 4 and 5 was in fact composed of complete mimosets 4.3, 4.4, 4.5 and 5.1. There is little to no sequence

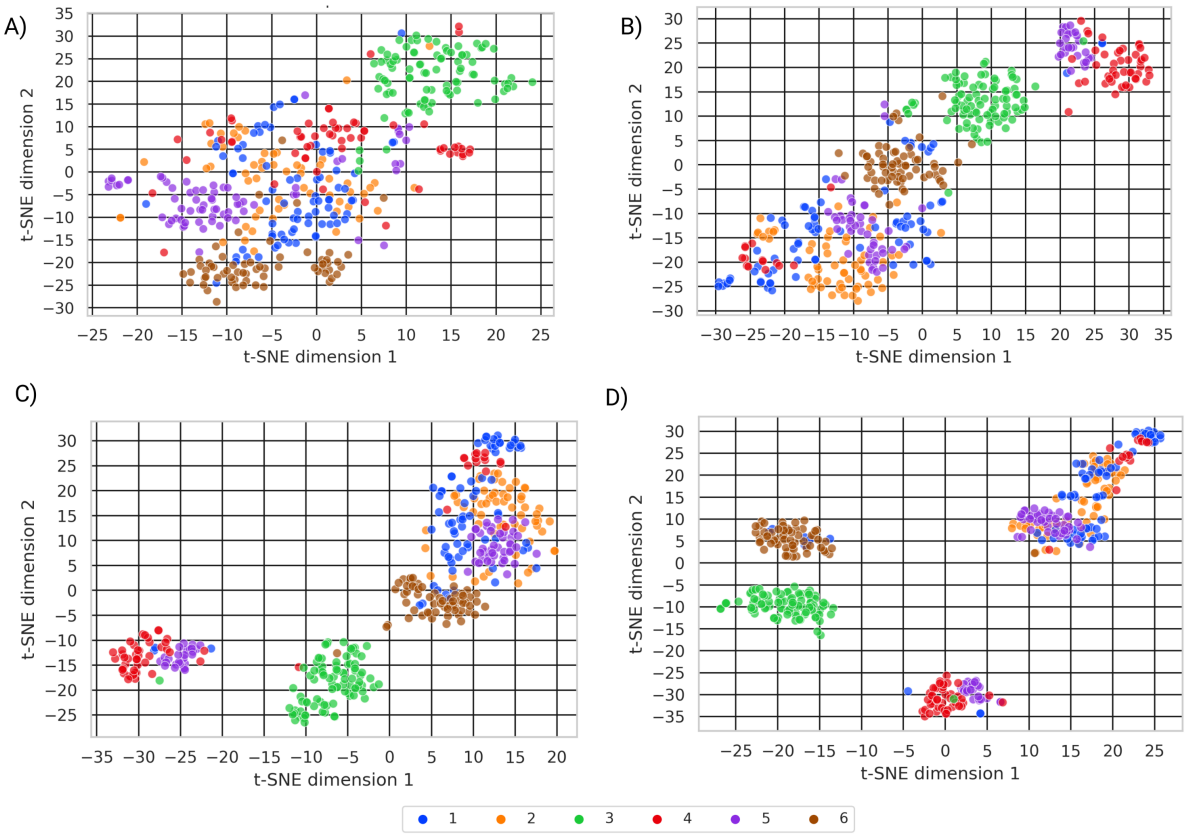

Figure 8. Visualization of dimensionality reduced peptide representations. Representations were generated from the Dataset A using following methods: A) ProtVec, B) UniRep c_final layer, C) BioVec, D) RGN_f2 layer. Peptides are color coded based on the binding target.

similarity shared between these mimosets; however, the peptide sequence length is identical between them. This led to a hypothesis that sequence length plays a role in grouping peptides together, which would constitute a bias in the dataset as it emphasises a feature which is not based on shared sequence properties.

Interestingly, the well separated group formed from 3.1 mimoset sequences shared not only the strong motif, but also constituted from peptides that had sequence length of 9 residues. No other mimoset in the Dataset A contains peptides with such length, thus possibly, the unique length may have contributed to such well-defined group. Target 6 group was formed from 6.1, 6.2 and 1.6 mimosets. All peptides in the group shared the SDL motif and the peptide length, however the degree to which either peptide length or shared motif contributed to mimoset grouping is unknown.

3.2.2 Dataset B

Since in Dataset A it was observed that the length of peptides in a mimoset can impact generated representations, the next step was to construct a new dataset which would contain standardised data, i.e. peptide length, phage library type and diversity would be the equal across mimosets to avoid possible external biases. Therefore, Dataset B was constructed by selecting mimosets that have been derived from Ph.D-12 phage library. Top 6 mAb targets with most peptide sequences were selected, in total 195 peptides were collected from 8 mimosets. Summary of Dataset B is given in Table 8, logos and details of libraries are given in Appendix Figure 20 and Appendix Table 22.

Table 8. Mimoset description across mAb targets in Dataset B. Motifs are based on MSA logos.

Target	Mimoset	Number of unique peptides, n	Panning Cycle	Library type	Peptide length	Motif
1	1	45	3	Linear	12	DKW
2	1	39	3	Linear	12	STSSxL
3	1	33	4	Linear	12	-
4	1	15	4	Linear	12	DxSTR
4	2	12	5	Linear	12	DxSTR
5	1	27	5	Linear	12	DxxP
6	1	19	3	Linear	12	-
6	2	5	4	Linear	12	PxxP

Similarly as on Dataset A, hierarchical clustering was also attempted on the Dataset B (see Figure 9). Cutoff value of 7 clusters was selected, minimising number of clusters while maximising proportion of peptides from single target in a cluster. Peptide distribution across clusters was analyzed in Table 9, and detailed peptide distribution per mimoset is given in Appendix Table 22. Peptides from targets 1, 2, 4 and 5 cluster well, i.e. target specific peptides are allocated in separate clusters while also mainly being only from one target. Peptides from target 3 are spread over several clusters. Analysing cluster 5 where target 3, 5 and 6 peptides are intermixing, it was found that 3.1 mimoset sequences shared WW motif, while sequences from other mimosets did not share explicit motifs. Similarly, cluster 4 contained mimoset sequences that did not share common sequence motifs. Overall hierarchical clustering performance on the Dataset B

was better than on the Dataset A. Dataset B contained 4 well-defined target specific peptide clusters, where each contained more than 85% of target specific peptides; Dataset A had only 2 of such clusters. Improvement in clustering can be attributed to stronger motifs present in target mimosets and possibly also to the standardised library used across panning experiments. Presence of motifs in most of the mimosets translated also into higher KNN sequence alignment baseline accuracy, ranging from 89% to 90% if compared to the Dataset A (see Table 10).

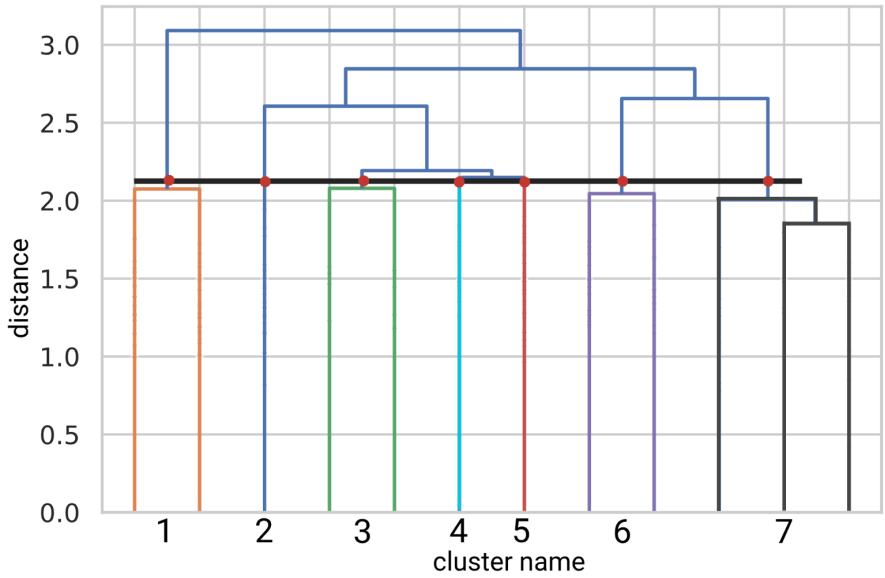

Figure 9. Hierarchical clustering dendrogram. Hierarchical clustering of peptides in Dataset B based on sequence pairwise alignment score. The resulting dendrogram is split into seven clusters. Vertical axis represents distance between clusters, horizontal axis represents assigned cluster. Clusters are color-coded.

Table 9. Target binding peptide distribution in fractions across different clusters in Dataset B.

	Cluster	1	2	3	4	5	6	7	Total peptides, n
Target									
1		1.00							45
2				0.05		0.05	0.87	0.03	39
3				0.24	0.39	0.33	0.03		33
4								1.00	27
5			0.85	0.04	0.07	0.04			27
6				0.04	0.17	0.79			24

Computational representations of peptide sequences in Dataset B were generated. Representation dimensions were reduced to two by t-SNE for visualization. Figure 10 contains selection of visualized representations from various methods, additional visualization can be seen in Appendix

Table 10. Weighted average accuracy values for KNN algorithm with different number of neighbours when predicting on Dataset B. KNN distance function was based on sequence alignment scores. Accuracy values were calculated using leave one out cross-validation.

	KNN number of neighbours, k				
	1	3	5	10	15
Alignment baseline	0.89	0.89	0.90	0.89	0.89

21. Compared to the Dataset A, the groups are much more diluted and are overlapping with each other, indicating that representations in high-dimensional space are not as well grouped as in Dataset A. However, marginal grouping still can be observed especially for the peptide mimosets that exhibit a motif - 1.1, 2.1, 5.1. Interestingly, Unirep c_final and RGN_f2 layer representations can group target 3 peptides which do not exhibit a motif, indicating that these representations can capture some common biological properties in that mimoset.

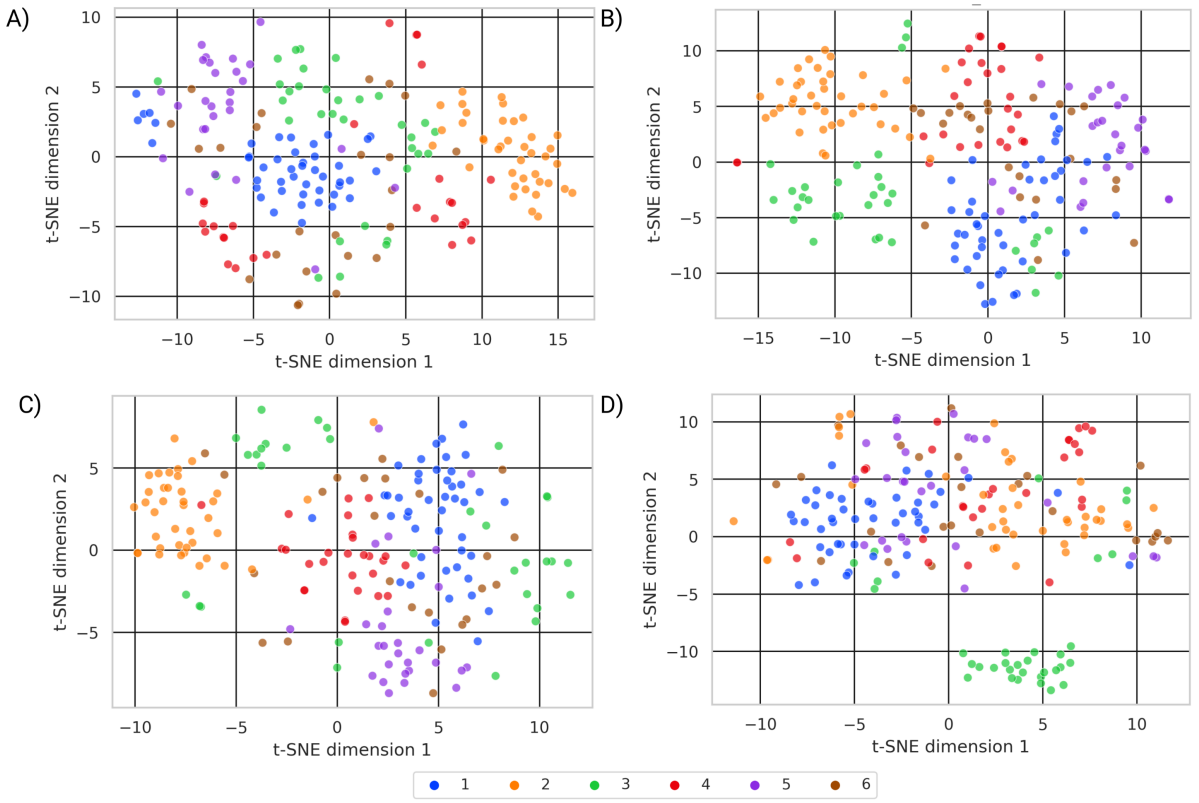

Figure 10. Visualization of dimensionality reduced peptide representations. Representations were generated: A) ProtVec, B) UniRep c_final layer, C) BioVec, D) RGN_f2 layer. Peptides are color coded based on the binding target.

KNN was applied to both full and dimension reduced sequence representations to evaluate the classification performance (results are summarized in Table 11). Both full and dimension reduced

Table 11. Weighted average accuracy values for KNN algorithm with different number of neighbours when predicting on different types of representations. Accuracy values were calculated using leave one out cross-validation. Representations were generated on Dataset B.

	Full dimensionality					Dimensionality reduced				
	KNN number of neighbours, k					KNN number of neighbours, k				
Representation	1	3	5	10	15	1	3	5	10	15
ProtVec	0.75	0.73	0.76	0.82	**0.82**	0.75	0.75	0.74	0.72	0.69
BioVec	0.73	0.80	0.71	0.79	0.78	0.75	0.71	0.74	0.73	0.69
UniRep c_final	**0.84**	**0.82**	0.82	**0.83**	0.80	**0.82**	**0.80**	**0.79**	0.72	0.71
UniRep h_final	0.65	0.63	0.62	0.61	0.59	0.65	0.61	0.63	0.62	0.62
UniRep h_avg	0.62	0.65	0.67	0.68	0.67	0.61	0.54	0.59	0.56	0.51
SeqVec	0.80	0.80	**0.83**	0.75	0.80	0.81	0.77	0.77	**0.76**	**0.76**
RGN_f1	0.71	0.76	0.80	0.80	0.78	0.74	0.75	0.76	0.74	0.71
RGN_f2	0.72	0.71	0.74	0.67	0.69	0.69	0.60	0.62	0.58	0.50
RGN_r1	0.58	0.64	0.60	0.68	0.70	0.58	0.58	0.52	0.54	0.48
RGN_r2	0.55	0.64	0.64	0.68	0.66	0.55	0.61	0.62	0.62	0.61
BERT	0.65	0.61	0.62	0.62	0.58	0.67	0.58	0.59	0.54	0.51
PLUS	0.40	0.48	0.42	0.44	0.40	0.38	0.40	0.45	0.38	0.36
Naive baseline	0.16	0.19	0.16	0.16	0.17	0.16	0.19	0.16	0.16	0.17

UniRep c_final and SeqVec representations perform on average better than other representation methods. Yet none of the representation methods led to KNN performance above the established alignment baseline performance. However, they lead to much better performance than the Naive baseline. By further investigating the classification performance on different targets across mimosets, it was found that target 6 peptides are the most difficult to classify accurately. Such result correlates with generated t-SNE plots, where target 6 representations do not form distinct groups. In contrast, target 1 and 2 were easier for KNN to classify, while classification of peptides from target 3, 4, 5 mimoset heavily depended on how well the representation algorithm could extract the relevant information.

By comparing KNN performance between Dataset A and Dataset B it could be seen that KNN performance was marginally worse for the best performing UniRep representations. However, other representation methods RGN, BERT, PLUS led to significantly worse KNN performance, possibly due to their representations containing sequence length bias.

3.2.3 Biopanning experiment

The amount of biopanning data deposited in the BDB is limited to very small mimosets, which were obtained under varying experimental conditions and with a variety of phage libraries. Recognising the need for more data obtained under standardised conditions in this area, we designed a standardised and controlled data acquisition pipeline in order to obtain a large scale dataset of peptide sequences that bind specific monoclonal antibodies (mAbs).

We selected three mAbs as panning targets, and we additionally included control target (BSA) that would allow to filter non-specific binding peptides. To obtain peptides that are specific for Anti–p53 antibody PAB 240 (PAB240), Anti–SARS–CoV–2 Spike glycoprotein S1 antibody CR3022 (CR3022) and Anti–Influenza A H1N1 hemagglutinin antibody C102 (H1N1). 12-mer peptide phage library was panned against the targets using surface panning procedure.

To validate whenever the binding phages were enriched by the assay, binding phage eluate was titered after each panning round. Titering results revealed enrichment of binding phages after each subsequent panning round, depicted in Table 12. The number of eluted phages panned against mAbs increased more than 100-fold after the second panning and more than 1000-fold after third panning round when compared to the first panning titers. Titering results demonstrate that through affinity selection rounds the enrichment of phages binding to some component of the assay took place, with later rounds of panning containing phages presumably displaying peptides with higher affinity to the target. Higher enrichment in non-control experiments was most likely due to enrichment of target specific peptides. The amount of enriched phages in CR3022 mAb sample was very close to the number of phages enriched by the BSA control, which may indicate that no phages specific to the target were enriched or that amount of binders enriched were low.

Table 12. Phage eluate titering results.

	mAb conc. (µg/mL)	Tween-20 (%)	CR3022 (pfu/µl)	H1N1 (pfu/µl)	pAB240 (pfu/µl)	BSA (pfu/µl)
Panning round 1	100	0.1	$2.65\text{x}10^{2}$	$1.90\text{x}10^{3}$	$1.98\text{x}10^{4}$	$2.76\text{x}10^{3}$
Panning round 2	66	0.3	$4.45\text{x}10^{4}$	$5.20\text{x}10^{5}$	$1.85\text{x}10^{7}$	$1.42\text{x}10^{4}$
Panning round 3	33	0.5	$8.80\text{x}10^{5}$	$1.05\text{x}10^{7}$	$6.50\text{x}10^{7}$	$3.80\text{x}10^{5}$

Peptide sequences binding the monoclonal antibodies were obtained by targeted sequencing of the phage genomes. In contrast to the the commonly used Sanger sequencing, NGS was used in order to generate data in a higher magnitude. In total, 11.2 million sequences were obtained from sequencing of 14 samples. After pre-processing steps, the total number of sequences that could be analysed was reduced to 7 million (see Appendix Table 17 for detailed information about sequence abundances in each pre-processing steps). The phage library used in the experiment was listed by manufacturer to be of diversity 10^9, and we used approximately 10^9 phages from Naive non-amplified library for sequencing, which if sequenced completely should yield approximately one clone per sequence. However, in this run sample sequencing depth was only 2.3x10^6, resulting in expected abundance value per sequence in the sample to be $2.3 \cdot 10^6/10^9 = 0.0023$. According to Poisson distribution this should result in sample containing 99.8% single clone sequences, 0.1% of duplicates and 0.1% of three or more clones. However, sequencing results of the Naive library did show that only 91.7% of sequences resulted in single clones, 7.3% were duplicates and 1% had 3 or more occurrences in some cases reaching 44 clones (Figure 11). Such abundance of clone sequences could indicate presence of parasitic sequences that skew population distribution due to rapid amplification. To quantify the extent to which such skew happens, 10^7 phages were amplified. Amplification with 10^9 phages was avoided to not introduce any additional biases from infecting at high MOI. Sequencing Naive amplified phage library lead to $2.0 \cdot 10^6$ reads. In case no amplification bias was present, result in sample should have followed the same Poisson distribution and contain 82% single repeats and 8.8% 2-5 repeats. However, only 18.4% of sequences resulted in the single peptide repeats, 2-5 times repeated sequences occupied 30% of the sequence pool, while single peptide sequence GENLMSVGLLRT occupied 2.7% of the whole sequencing pool. Such skew in abundances towards some select pool of peptides after bacterial amplification resulted in heavily unbalanced phage population. In case of biopanning experiments that involve multiple rounds of selection and amplification, such quickly amplifying parasitic sequences lead to resulting sample containing quickly-amplifying rather than tightly-binding phage sequences.

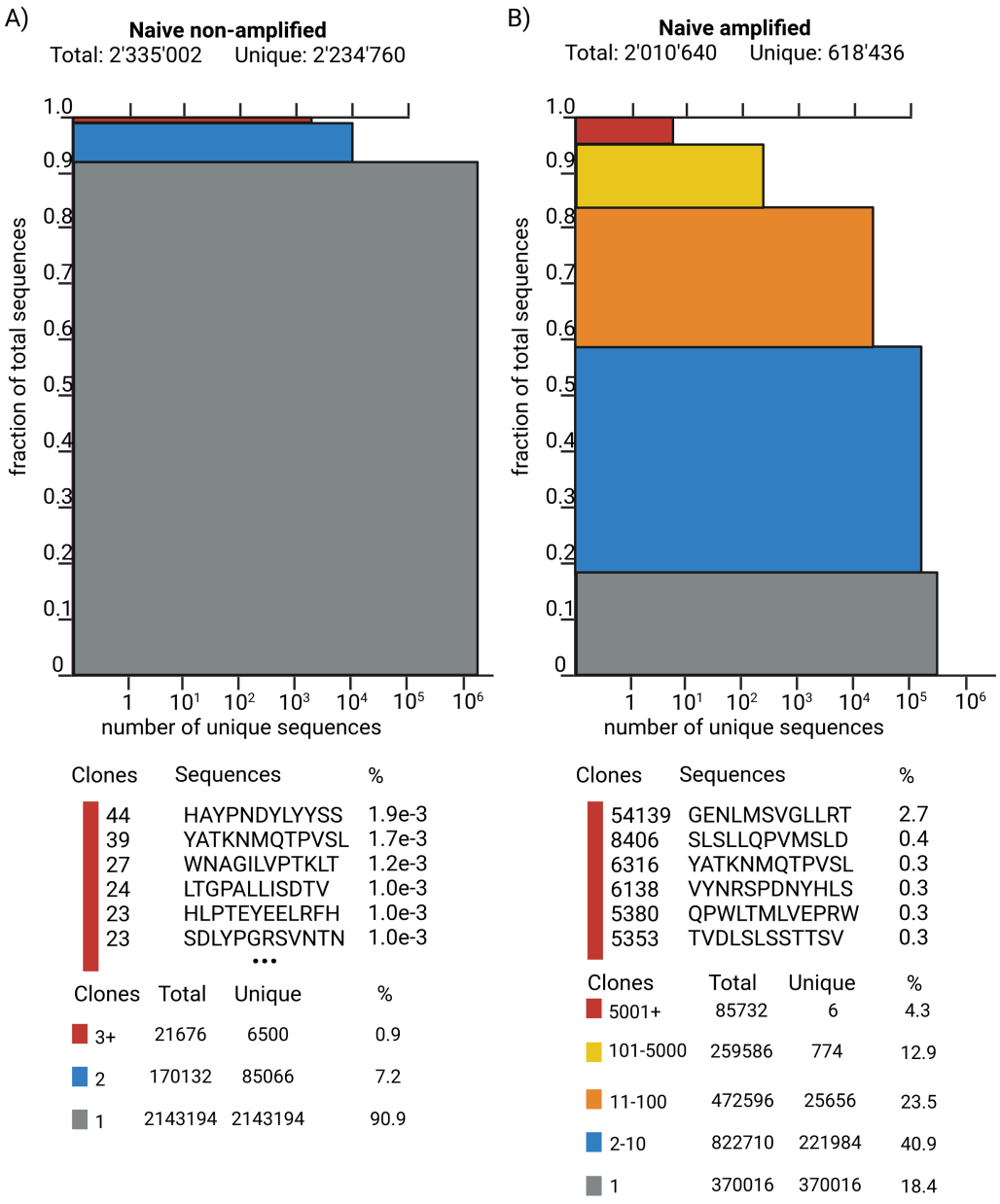

Figure 11. Stacked bar plot of peptide sequence abundances in Naive and Naive amplified libraries. Vertical axis shows percentage of total sequences that are occupied by particular clones, horizontal axis shows the amount of unique sequences that are present among the clones.

In total, 12 samples were obtained from three-round affinity selection experiment against 3 mAbs and BSA control. Additional pre-processing steps were carried out for the monoclonal antibody target samples, removing sequences that are present in BSA control sample and removing sequences that are present in samples panned against other mAb targets. Stacked bar plots for each sample were generated to analyse enrichment of sequences (Appendix Figures 22, 23, 24, 25). After each panning round the proportion of unique sequences decreased, while the quantities for various clone sequences increased several fold, thus showing possible enrichment towards some set of binders. In all three samples, the top enriched peptides after second biopanning round could be observed as top peptides in the third round. Interestingly, in PAB240 sample after third panning round the sequence TLHKVVLRSAIP occupied 41% from total sequences, indicating very strong enrichment. However, this sequence was also among top 20 sequences with the most clones in the Naive amplified library indicating that enrichment could be linked to rate of amplification. Therefore, before conducting any additional analysis on the most enriched sequences, it was important to mitigate potential amplification bias.

To mitigate the effect of quickly-amplifying sequences, sequencing data was normalized using amplified Naive library sequencing data. The normalization penalized those sequences in the sample that were over-abundant in the Naive amplified sample, as a result such quickly amplifying sequences would not be present as top enriched ones. This allowed to rank sequences by their normalized clone count and prioritize sequences that have been enriched by their binding properties, rather than rate of amplification. Normalized stacked bar plots were generated to visualize the distribution and the enrichment towards target binding sequences (Figures 12,13,14,15). Analysis of the normalized data accentuated, how the enrichment of specific binders emerged already in the second panning round. The data from the third panning round returned the same top sequences, albeit in slightly different order. Comparing the top 100 enriched sequences before and after normalization revealed that on average there were 10 sequences that also were in top 100 enriched sequences in the Naive amplified sample. The normalization process heavily penalized such sequences, removing them from top enriched sequences, and minimized the amplification bias influence on potential binder selection.

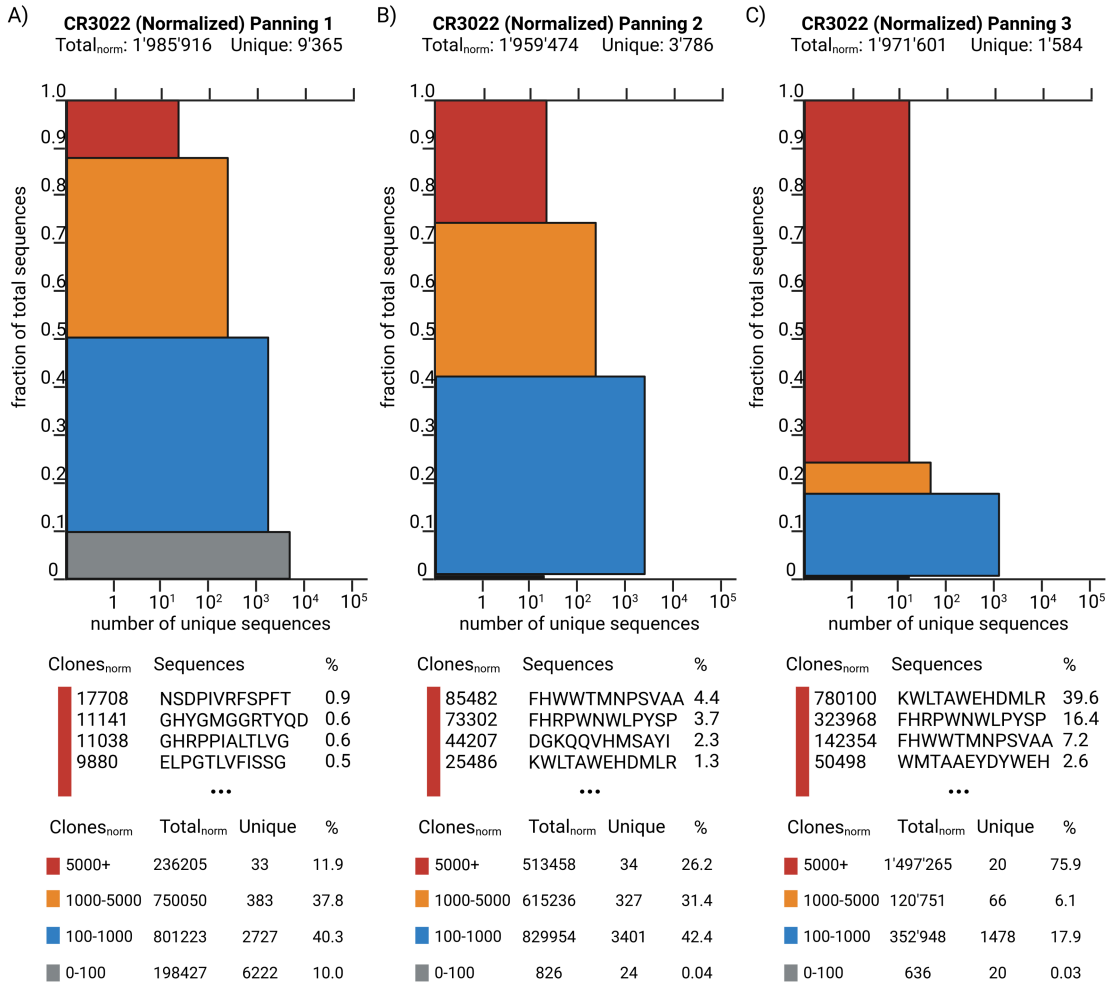

Figure 12. Stacked bar plot of pre-processed normalized sequencing data from CR3022 mAb sample. Y-axis shows the fraction from total n_{norm} that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total n_{norm}.

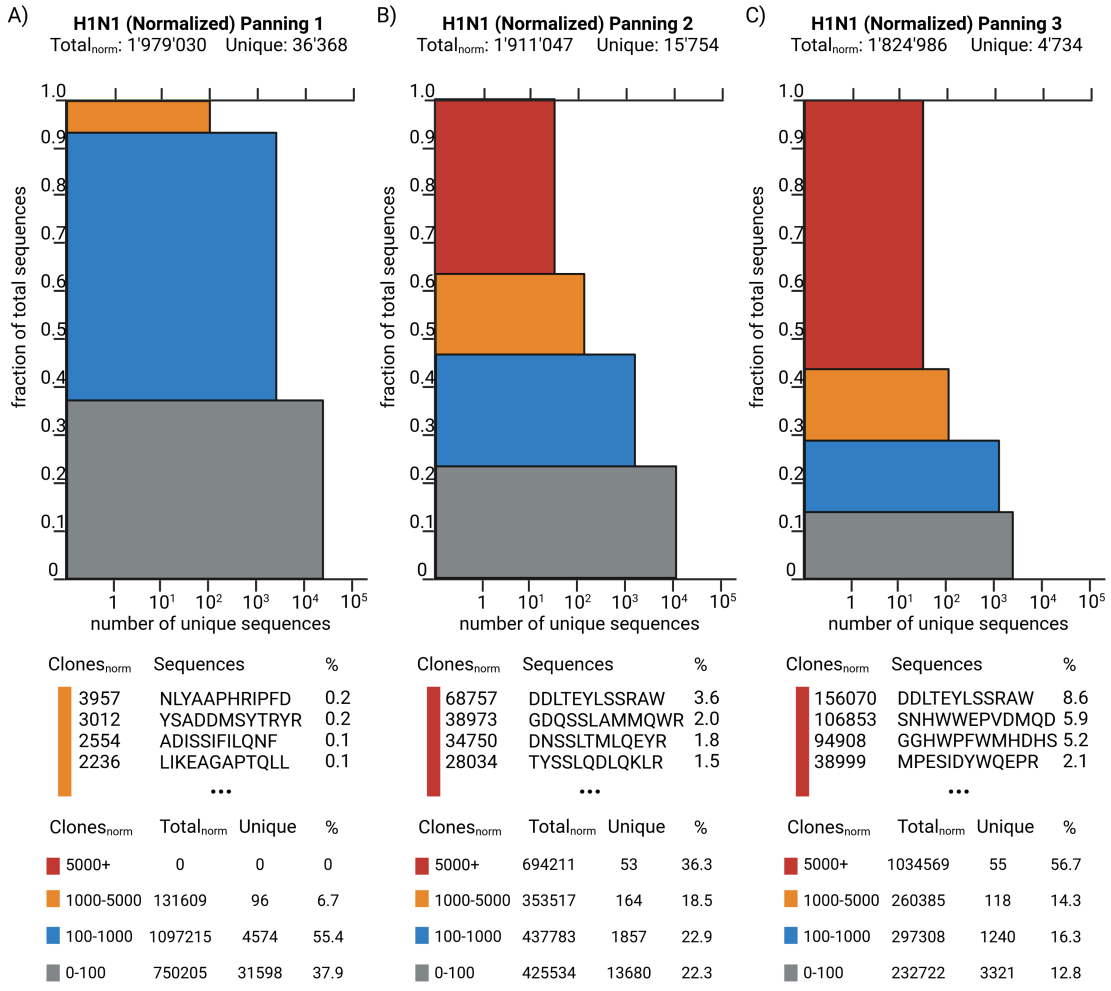

Figure 13. Stacked bar plot of pre-processed normalized sequencing data from H1N1 mAb sample. Y-axis shows the fraction from total n_{norm} that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total n_{norm}.

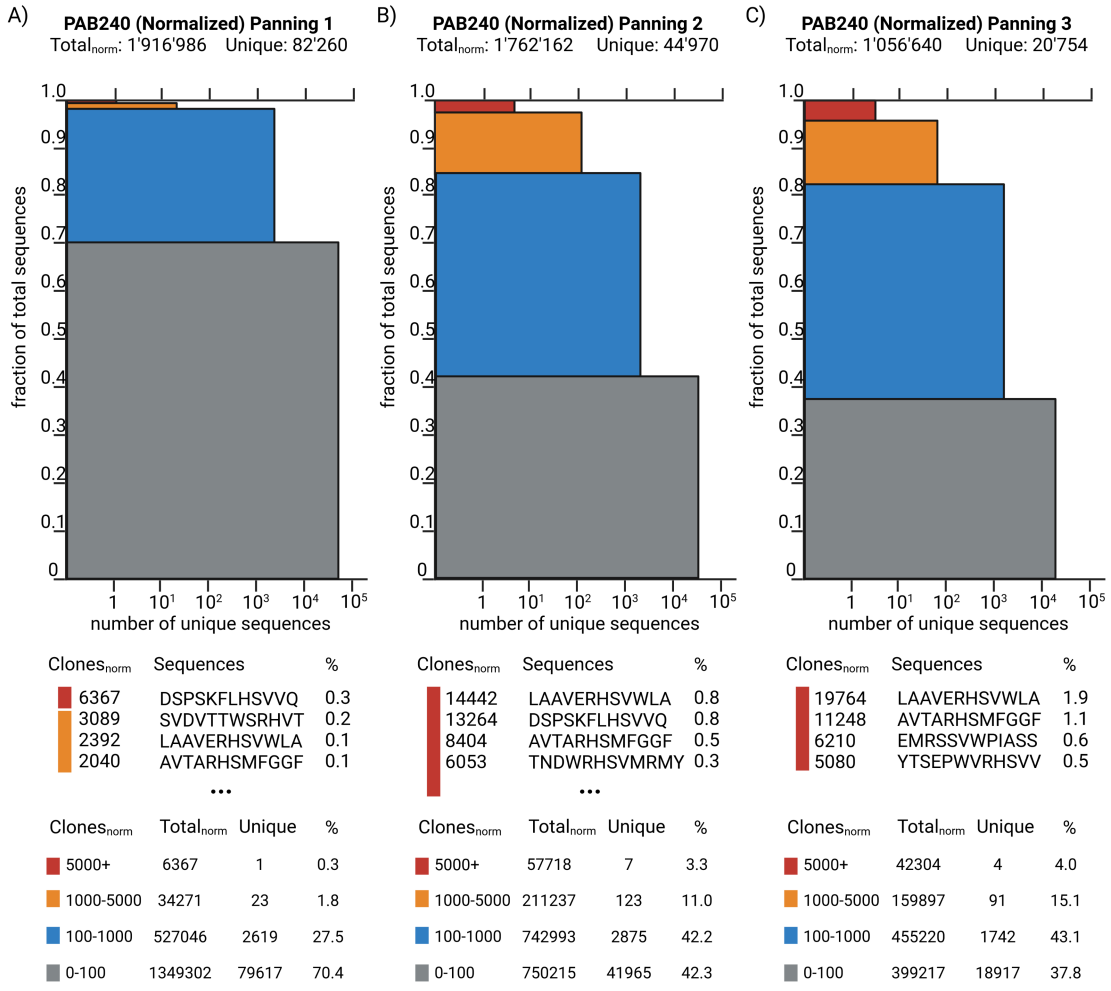

Figure 14. Stacked bar plot of pre-processed normalized sequencing data from PAB240 mAb sample. Y-axis shows the fraction from total n_{norm} that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total n_{norm}.

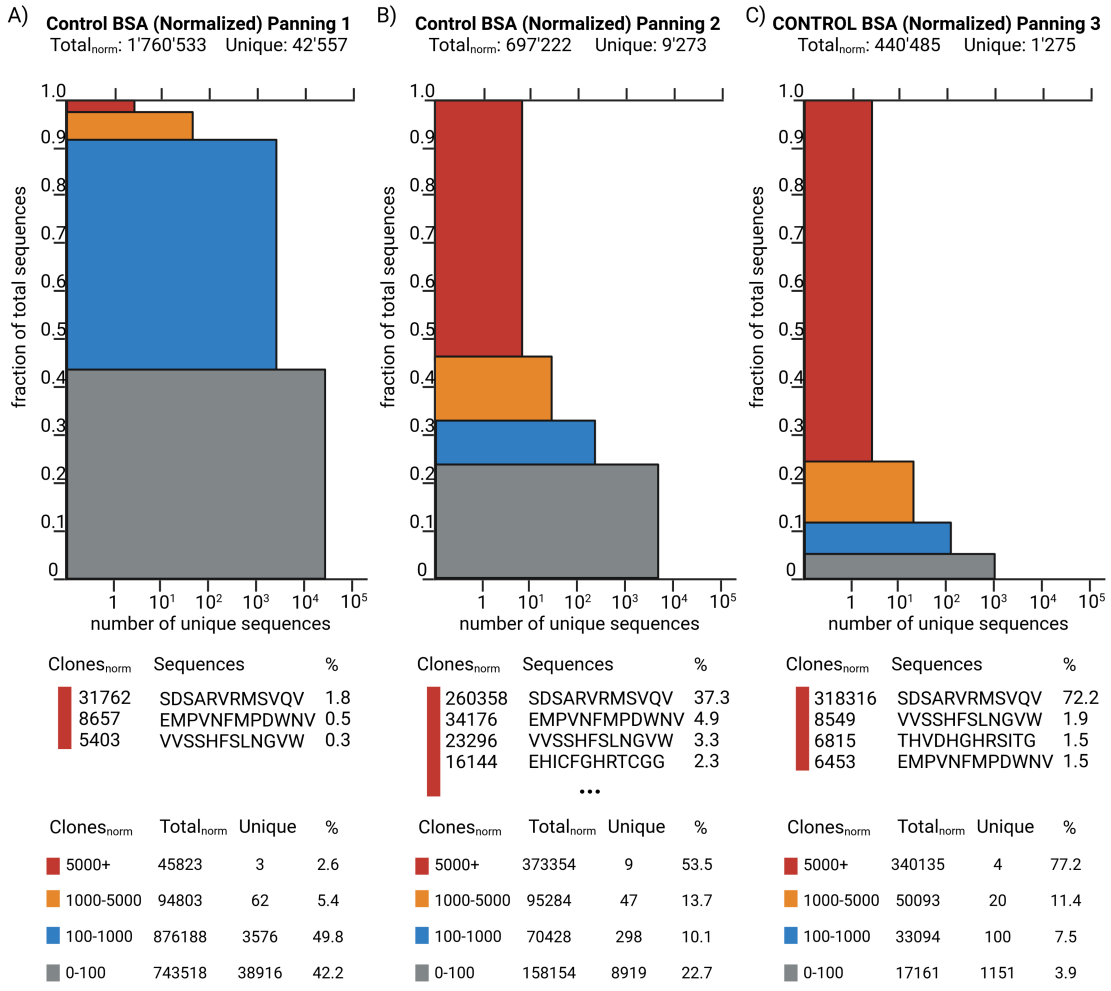

Figure 15. Stacked bar plot of pre-processed normalized sequencing data from control (BSA) sample. Y-axis shows the fraction from total n_{norm} that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total n_{norm}.

In order to highlight any shared motifs for each mAb, the top 100 enriched normalized sequences for each mAb target were used with MEME motif extraction algorithm. The MEME extracted logos for mimosets that exhibited a motif are depicted in Figure 16. Normalized panning results against PAB240 revealed strong motif RHS[V/M][V/L/I] already after the first round with 94 sequences out of 100 exhibiting this motif. In second and third panning 100 and 99 exhibiting this motif, respectively. Motif SSLxx[M/L]Q was present in 59 and 39 H1N1 binding sequences from panning round 2 and 3 respectively - first panning round did not return this motif. Interestingly, no motif emerged from the samples panned against the CR3022 target, despite the enrichment of certain set of peptides over each panning round.

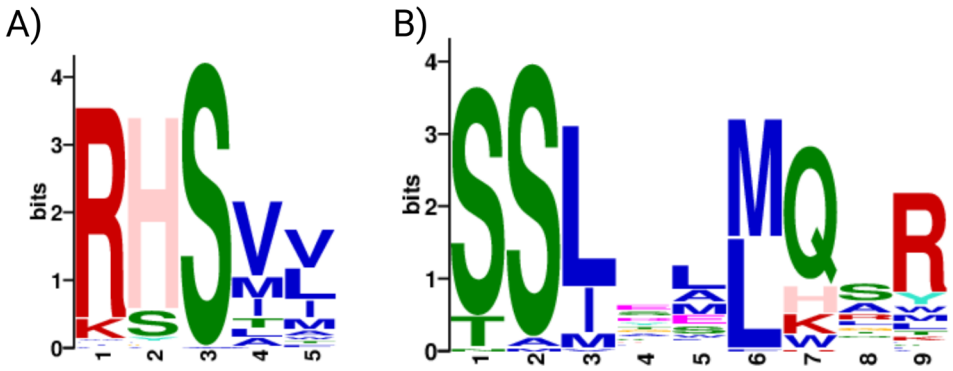

Figure 16. MEME motif logos. A) PAB240 motif Logo generated on the normalized top 100 sequences from the first panning round. B) H1N1 motif Logo generated on the normalized top 100 sequences from the second panning round. Horizontal axis depict alignment position while vertical axis show the information content in bits.

3.2.4 Representation performance on experimental data

The newly generated rich mimoset datasets that were obtained under standardised conditions and vigorously pre-processed, were assessed using computational techniques as described in sections 3.2.1 and 3.2.2. We took top 50 enriched sequences from normalized mAbs samples from the third panning and generated their full and dimensionality reduced sequence representations. Naive and sequence alignment KNN baselines were generated and compared to KNN performance on sequence representations. From results in Table 13 it can be seen the UniRep c_final representations result in the best performance on average. Interestingly, the KNN performance on UniRep is by approx. 10% lower than the alignment baseline, thus confirming the observation that the performance of representation methods declines substantially when applied to a properly

controlled dataset. The sequence alignment baseline in Table 14 was significantly lower than in the Dataset B, indicating a much larger sequence diversity present in the experimentally generated data.

Table 13. KNN weighted accuracy values on experimental data.

	Full dimensionality					Dimensionality reduced				
	KNN number of neighbours, k					KNN number of neighbours, k				
	1	3	5	10	15	1	3	5	10	15
ProtVec	0.63	0.65	0.65	0.60	0.60	0.59	0.61	0.63	0.54	0.57
BioVec	0.59	0.62	0.54	0.59	0.53	0.59	0.61	0.63	0.54	0.57
UniRep c_final	**0.68**	**0.70**	**0.73**	**0.68**	0.67	**0.68**	**0.66**	0.65	**0.67**	**0.62**
UniRep h_final	0.54	0.49	0.52	0.57	0.54	0.51	0.51	0.43	0.42	0.41
UniRep h_avg	0.62	0.65	0.67	0.66	0.64	0.59	0.64	**0.66**	0.63	0.62
SeqVec	0.66	0.65	0.67	0.63	0.61	0.58	0.63	0.54	0.57	0.59
RGN_f1	0.59	0.63	0.63	0.65	0.65	0.59	0.61	0.63	0.61	0.61
RGN_f2	0.57	0.64	0.64	0.61	0.58	0.58	0.63	0.54	0.54	0.50
RGN_r1	0.62	0.66	0.61	0.65	**0.69**	0.63	0.59	0.64	0.56	0.53
RGN_r2	0.50	0.53	0.48	0.52	0.54	0.48	0.52	0.49	0.48	0.50
BERT	0.52	0.61	0.55	0.59	0.57	0.55	0.54	0.55	0.49	0.47
PLUS	0.45	0.47	0.40	0.43	0.39	0.46	0.38	0.42	0.47	0.47
Naive baseline	0.37	0.32	0.39	0.37	0.30	0.37	0.32	0.39	0.37	0.30

Table 14. KNN weighted accuracy values for sequence alignment baseline generated on experimental data.

	KNN number of neighbours, k				
	1	3	5	10	15
Alignment baseline	0.81	0.78	0.81	0.80	0.77

Dimensionality reduced representations were plotted in Figure 17, other representation method plots are available in Appendix 26. From plots it was observed that representations generated on sequences with strong motif (PAB240) were forming a group (green dots). Representations generated on samples with mostly diverse sequences not sharing strong motifs (CR3022 and H1N1) resulted in either no distinct group formation or small sub-group formation. Analysing formed subgroups in UniRep c_final representation plot it was found that the subgroup formed by 13 CR3022 sequences shared WW residue pattern, which might be relevant for binding. Similarly, the subgroup formed from 15 H1N1 peptide sequences shared SSL motif. The poor formation of groups in t-SNE plots showed that from implemented representation techniques none can reliably group target specific binders, even if such binders share a strong motif as it was in case of PAB240.

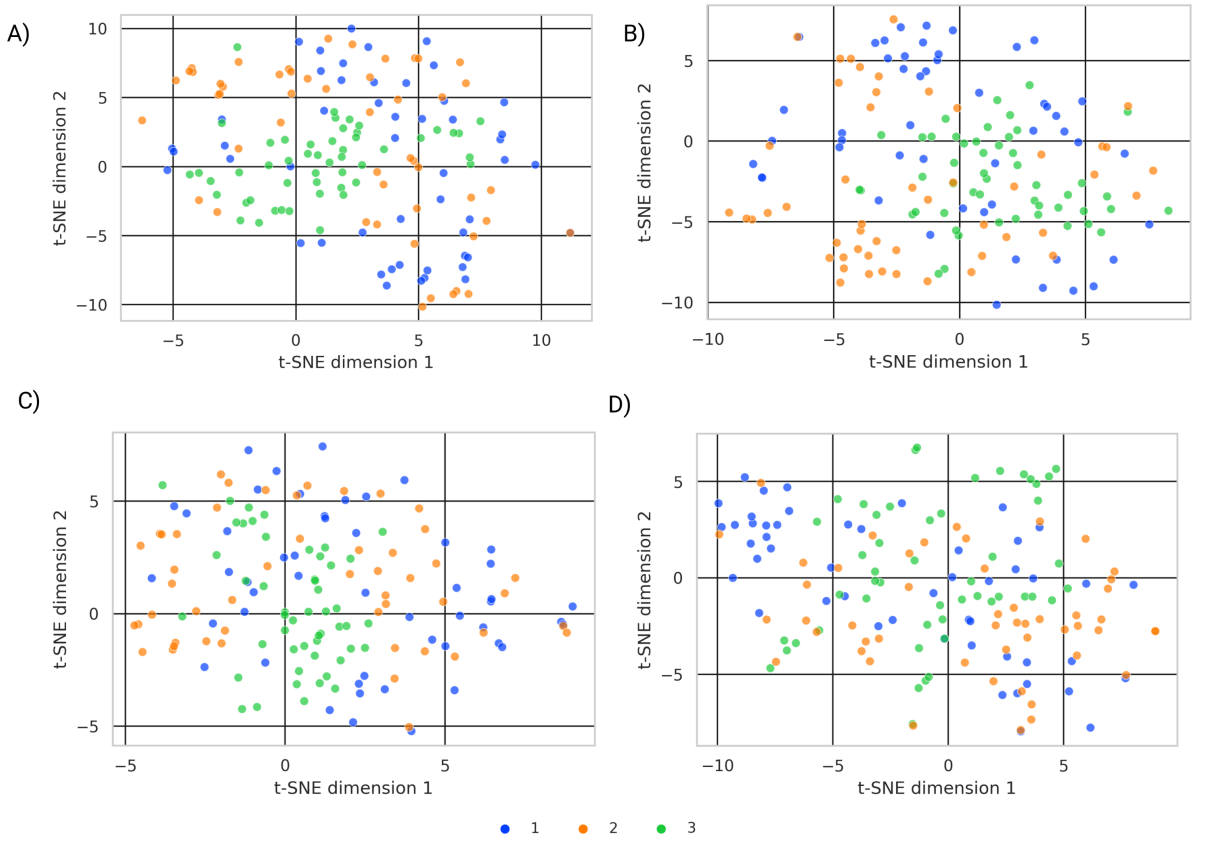

Figure 17. Visualization of dimensionality reduced peptide representations. Representations were generated: A) ProtVec, B) UniRep c_final layer, C) BioVec, D) RGN_f2 layer. Peptides are color coded based on the binding target. Legend; 1 - CR3022, 2 - H1N1, 3- PAB240.

3.3 DISCUSSION

Our initial assumption that target specific binding peptides share a strict motif was proven wrong when we attempted to construct a sequence alignment baseline based on hierarchical clustering. Intermixing of peptides targeting different mAbs in a cluster and separation of motif-shared mimosets across different clusters indicated that strict clustering by sequence similarity is non-viable. It might be the case that together clustered sequences would still bind the target, however, data source did not contain such cross-targeting information and testing this possibility by experimental means was out of scope for this work. Thus we decided to relax this assumption, and instead of expecting to obtain single clusters that group the majority of peptides binding a specific target together, we switched to using KNN as the method to group related peptides together. The new assumption underlying this choice was that there might be multiple groups of related sequences that are dissimilar group-to-group but which all bind to the same target antibody.

Visualization of t-SNE dimensionality reduced protein representations on Dataset A revealed that some representations suffer from bias induced by the peptide length in the mimoset. As a result, the mimosets are not grouped by their biological properties relevant for binding but rather by shared sequence length. Such bias exploitation may be why the KNN performance on the BioVec and UniRep c_final representations marginally outperform the sequence alignment baseline KNN performance on the Dataset A. The degree to which sequence length bias affects representations remains an open question. We minimized this bias by constructing a standardised Dataset B. Re-evaluating the representations on the standardised dataset revealed the KNN performance on some of the methods (BioVec, PLUS, RGN_r1, RGN_r2, UniRep h_avg) deteriorated substantially, indicating their sensitivity towards sequence length. Also, KNN performance on other representation methods deteriorated, albeit to a lesser extent.

We have found that commonly used deep learning based protein sequence representation methods, when applied to a standardised dataset of short-peptide sequences, are inferior to sequence alignment baseline for mapping epitope landscape using the KNN algorithm. Although sequence alignment baseline performance was negatively affected by the lack of shared motifs across se-

quences in target-specific mimosets, the representational models still fell short on outperforming it. Part of the issue might be that these models have been pre-trained on full-length protein sequences, thus resulting in suboptimal representations generated on short peptide sequences. The second part of the issue might underlie the pre-training task that lacks incorporation of information that is relevant to the binding - structural and physicochemical information. Although RGN embeddings have been structurally pre-trained, they did not outperform the baseline. Our results indicate that computational methods need additional fine-tuning oriented on short-peptides and use pre-training task or a combination of tasks that incorporate information relevant to binding paratope.

Future improvements of data-driven models in this field require the availability of large-scale data, especially for deep learning based method training. To address the issue of mimotope data scarcity, we developed and implemented the method for data acquisition using the peptide phage display technique combined with the NGS.

Sequencing results of Naive phage libraries showed the magnitude of amplification bias existing in our library; such bias has been previously described (Matochko et al., 2013). We mitigated the effect using the normalization technique described by the work of Juds et al. (2020). Several additional techniques have been described in the literature to minimize amplification bias further: filtering data using public repositories of quickly-amplifying sequences (Pleiko et al., 2020), amplifying phages in semi-solid media or in emulsion droplets (Pleiko et al., 2020; Derda et al., 2010). Future implementations of the established pipeline might utilize some of the mentioned additional techniques.

MEME motif analysis from normalized enriched peptide sequences against Anti-p53 PAB240 mAb revealed that it is possible to retrieve the consensus target binding motif already after the first panning round. We believe that such result was possible mainly due to the strong enrichment of this motif and the high abundance of refined sequence data provided by NGS. This motif has already been discovered previously by phage display after three rounds of selection coupled with Sanger sequencing (Stephen et al., 1995). The replicability of previously published results confirm the robustness of the NGS-based approach and outline how NG-sequencing depth allows

to unravel existing motifs among binding peptides already after the first panning round, while typical phage display experiments coupled with Sanger sequencing require several rounds of panning before motifs can be observed.
The motif search of Anti-influenza A H1N1 mAb mimotope revealed the presence of the previously unknown motif (SSLxx[M/L]Q) after the second and third panning. Previous study results panning against this mAb did not result in retrieving any motif (Zhong et al., 2011), possibly due to low sequencing depth from Sanger sequencing.

Motif retrieval from Anti-SARS-CoV-2 (CR3022) mAb mimotope was unsuccessful. We speculate that there might be two mechanisms or the interplay between them leading to this result. It is possible that epitope repertoire recognized by CR3022 antibody is highly diverse in terms of sequence; therefore, no common motif emerged in the top enriched normalized sequences. Such mimosets without definitive motif have been described in the literature and were also present in the datasets analyzed in this work (Zhong et al., 2011; Tarnovitski et al., 2006). Another possibility may be due to the antibody not binding the assay plate or detaching from it during washing steps. Such technical issues would lead to the loss of the target and the enriched sequence. We can speculate from our data that the latter mechanism might have likely produced such a result due to CR3022 sample titers being in the same magnitude as the BSA control titers and due to the significant overlap with sequences present in the BSA control sample (approx 80% of sequences overlapped between pre-filtered CR3022 and BSA control).

We believe that it may be the case that mimosets obtained from phage display experiments using Sanger sequencing could contain a considerable amount of quickly-amplifying parasitic sequences, especially if the experiments do not account for such sequences. Many enrichment rounds and amplification rounds would allow strong enrichment of those sequences (as in our PAB240 mAb sample). Thus when the phage population is selected for sequencing by randomly picking phage clones, there would be a high chance of picking a parasitic sequence. Without vigorous experimental validation and filtering of such non-specific sequences, the data needs to be treated with caution.

It should be emphasized that the findings of this work in relation to discovered motifs and enriched sequences through phage display biopanning require additional validation of antibody-peptide complex formation. Despite implementing vigorous pre-processing steps to refine antibody binding peptides, it is crucial that they are validated by experimental means. Methods that allow high-throughput characterisation of their affinity to the antibody are preferred, for example via microfluidic diffusional sizing (Arter et al., 2020).

All in all, the results in this work show that further improvements are needed for representation methods. The established experimental pipeline for obtaining large scale mimoset data addresses first step of improving such data-driven methods. Moreover, we believe that during this work we have also obtained data that allows us to evaluate which sequences were poorly enriched or were not enriched at all. We believe that such data may also be beneficial in future representation method improvement.

4 SUMMARY

We have found that pre-trained deep learning based sequence representations did not outperform the sequence alignment baseline for mapping antibody binding peptides. Although implemented representations led to better than Naive performance, they all suffered from some degree of peptide length induced bias.

Commonly used Sanger sequencing in tandem with phage display limits the amount of mimotope data obtained and may suffer from the presence of the amplification biased sequences. In this work, we have combined phage display with NGS and applied vigorous pre-processing steps, thus acquiring a high-quality sample. As a result, it was possible not only to confirm the anti-p53 mimoset binding motif with data in the literature but also to unravel the previously unknown mimotope motif possibly binding anti-influenza A H1N1 antibody. Therefore, these results validate the promising approach for deep panning monoclonal antibody mimotopes.

The COVID-19 pandemic has clearly exemplified how rapid vaccine design and implementation is an essential step in mitigating the damage novel pathogens exhibit. Considering the possibility of additional pathogen spill-over events to mankind, understanding the fundamental characteristics of antibody-antigen sterical and physicochemial interactions will be crucial for improved vaccine development pipelines. Currently nascent *in silico* methods for characterising possible immune escaping variants will play major role in the viral arms race.

Overall, this work's results outline the need to improve deep learning based sequence representations, specifically on short-peptide sequences. Additionally, the established methodology for acquiring large-scale high-quality parallel data of antibody mimotopes serves as a fundamental first step toward improving protein sequence-based computational models.

ACKNOWLEDGEMENTS

I want to thank my supervisors, Erik Abner and Alekszej Morgunov, for granting me the chance to work on this exciting topic and assisting with advice when I got stuck. Special thanks to Tuomas Knowles for giving me the privilege of research freedom and for hosting me in his research group.

REFERENCES

Aho, A. V. (1991). Algorithms for finding patterns in strings, handbook of theoretical computer science (vol. a): algorithms and complexity.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322.

AlQuraishi, M. (2018). Rgn protein sequence representations. `https://github.com/aqlaboratory/rgn`. Online available; accessed at 16/10/2021.

AlQuraishi, M. (2019a). End-to-End Differentiable Learning of Protein Structure. *Cell Systems*, 8(4):292–301.e3.

AlQuraishi, M. (2019b). Proteinnet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311.

Arter, W. E., Levin, A., Krainer, G., and Knowles, T. P. J. (2020). Microfluidic approaches for the analysis of protein–protein interactions in solution. *Biophysical Reviews*, 12(2):575–585.

Asgari, E. and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE*, 10(11):1–15.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue):W202–W208.

Bairoch, A. and Apweiler, R. (2000). The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48.

Barreto, K., Maruthachalam, B. V., Hill, W., Hogan, D., Sutherland, A. R., Kusalik, A., Fonge, H., DeCoteau, J. F., and Geyer, C. (2019). Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. *Nucleic Acids Research*, 47(9):e50–e50.

Bennett, N. J., Gagic, D., Sutherland-Smith, A. J., and Rakonjac, J. (2011). Characterization of a dual-function domain that mediates membrane insertion and excision of ff filamentous bacteriophage. *Journal of Molecular Biology*, 411(5):972–985.

Brammer, L. A., Bolduc, B., Kass, J. L., Felice, K. M., Noren, C. J., and Hall, M. F. (2008). A target-unrelated peptide in an m13 phage display library traced to an advantageous mutation in the gene ii ribosome-binding site. *Analytical Biochemistry*, 373(1):88–98.

Changgeon Lee, Woo Jun, K. S. (2017). Biovec protein sequence representations. `https://github.com/jowoojun/biovec`. Online available; accessed at 15/09/2021.

Chen, F., Jiang, R., Wang, Y., Zhu, M., Zhang, X., Dong, S., Shi, H., and Wang, L. (2017). Recombinant phage elicits protective immune response against systemic s. globosa infection in mouse model. *Scientific Reports*, 7(1):42024.

Clokie, M. R., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011). Phages in nature. *Bacteriophage*, 1(1):31–45.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Cunningham, P. and Delany, S. J. (2020). k-nearest neighbour classifiers: 2nd edition (with python examples). *CoRR*, abs/2004.04523.

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., Yang, K. K., Min, S., Yoon, S., Morton, J. T., and Rost, B. (2021). Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113.

Davidson, E. and Doranz, B. J. (2014). A high-throughput shotgun mutagenesis approach to mapping b-cell antibody epitopes. *Immunology*, 143(1):13–20.

Derda, R., Tang, S. K. Y., and Whitesides, G. M. (2010). Uniform amplification of phage with different growth characteristics in individual compartments consisting of monodisperse droplets. *Angewandte Chemie (International ed. in English)*, 49(31):5301–5304.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2020). Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *CoRR*, abs/2007.06225.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285.

F.R.S., K. P. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Fuh, G. and Sidhu, S. S. (2000). Efficient phage display of polypeptides fused to the carboxy-terminus of the m13 gene-3 minor coat protein. *FEBS Letters*, 480(2-3):231–234.

Gao, C., Mao, S., Lo, C.-H. L., Wirsching, P., Lerner, R. A., and Janda, K. D. (1999). Making artificial antibodies: A format for phage display of combinatorial heterodimeric arrays. *Proceedings of the National Academy of Sciences*, 96(11):6025–6030.

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic acids research*, 36(9):3025–3030.

He, B., Chai, G., Duan, Y., Yan, Z., Qiu, L., Zhang, H., Liu, Z., He, Q., Han, K., Ru, B., Guo, F.-B., Ding, H., Lin, H., Wang, X., Rao, N., Zhou, P., and Huang, J. (2015). BDB: biopanning data bank. *Nucleic Acids Research*, 44(D1):D1127–D1132.

He, B., Jiang, L., Duan, Y., Chai, G., Fang, Y., Kang, J., Yu, M., Li, N., Tang, Z., Yao, P., Wu, P., Derda, R., and Huang, J. (2018). Biopanning data bank 2018: hugging next generation phage display. *Database*, 2018. bay032.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723.

Henry, T. J. and Pratt, D. (1969). The proteins of bacteriophage m13. *Proceedings of the National Academy of Sciences of the United States of America*, 62(3):800–807.

Hurwitz, A. M., Huang, W., Estes, M. K., Atmar, R. L., and Palzkill, T. (2017). Deep sequencing of phage-displayed peptide libraries reveals sequence motif that detects norovirus. *Protein Engineering, Design and Selection*, 30(2):129–139.

Iannolo, G., Minenkova, O., Petruzzelli, R., and Cesareni, G. (1995). Modifying filamentous phage capsid: Limits in the size of the major capsid protein. *Journal of Molecular Biology*, 248(4):835–844.

Ibtehaz, N. and Kihara, D. (2021). Application of sequence embedding in protein sequence-based predictions.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer New York.

Jespers, L. S., Messens, J. H., Keyser, A. D., Eeckhout, D., Brande, I. V. D., Gansemans, Y. G., Lauwereys, M. J., Vlasuk, G. P., and Stanssens, P. E. (1995). Surface expression and ligand-based selection of cdnas fused to filamentous phage gene vi. *Bio/Technology*, 13(4):378–382.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). Scipy open source scientific tools for python. `http://www.scipy.org`. Online available; accessed at 1/05/2022.

Juds, C., Schmidt, J., Weller, M. G., Lange, T., Beck, U., Conrad, T., and Börner, H. G. (2020). Combining phage display and next-generation sequencing for materials sciences: A case study on probing polypropylene surfaces. *Journal of the American Chemical Society*, 142(24):10624–10628.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Karimi, M., Mirshekari, H., Moosavi Basri, S. M., Bahrami, S., Moghoofei, M., and Hamblin, M. R. (2016). Bacteriophages and phage-inspired nanocarriers for targeted delivery of therapeutic cargos. *Advanced drug delivery reviews*, 106(Pt A):45–62.

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., Lee, J.-I., Suh, Y.-L., Ku, B. M., Eum, H. H., Choi, S., Choi, Y.-L., Joung, J.-G., Park, W.-Y., Jung, H. A., Sun, J.-M., Lee, S.-H., Ahn, J. S., Park, K., Ahn, M.-J., and Lee, H.-O. (2020). Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications*, 11(1):2285.

Kimothi, D., Biyani, P., and Hogan, J. M. (2019). Sequence representations and their utility for predicting protein-protein interactions. *bioRxiv*.

Koide, A., Wojcik, J., Gilbreth, R. N., Reichel, A., Piehler, J., and Koide, S. (2009). Accelerating phage-display library selection by reversible and site-specific biotinylation. *Protein Engineering, Design and Selection*, 22(11):685–690.

Kyu Ko, S. W. (2016). Biovec protein sequence representations. `https://github.com/kyu999/biovec`. Online available; accessed at 10/09/2021.

Lever, J., Krzywinski, M., and Altman, N. (2017). Principal component analysis. *Nature Methods*, 14(7):641–642.

Loset, G. , Roos, N., Bogen, B., and Sandlie, I. (2011). Expanding the versatility of phage display ii: Improved affinity selection of folded domains on protein vii and ix of the filamentous phage. *PLOS ONE*, 6(2):1–10.

Lowman, H. (2013). Phage display for protein binding. In Lennarz, W. J. and Lane, M. D., editors, *Encyclopedia of Biological Chemistry (Second Edition)*, pages 431–436. Academic Press, Waltham, second edition edition.

Makky, S., Dawoud, A., Safwat, A., Abdelsattar, A. S., Rezk, N., and El-Shibiny, A. (2021). The bacteriophage decides own tracks: When they are with or against the bacteria. *Current Research in Microbial Sciences*, 2:100050.

Matochko, W. L., Chu, K., Jin, B., Lee, S. W., Whitesides, G. M., and Derda, R. (2012). Deep sequencing analysis of phage libraries using illumina platform. *Methods*, 58(1):47–55. Phage-Display and Related Areas.

Matochko, W. L., Cory Li, S., Tang, S. K., and Derda, R. (2013). Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Research*, 42(3):1784–1798.

McConnell, S. J., Dinh, T., Le, M.-H., and Spinella, D. G. (1999). Biopanning phage display libraries using magnetic beads vs. polystyrene plates. *BioTechniques*, 26(2):208–214.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Min, S., Park, S., Kim, S., Choi, H.-S., Lee, B., and Yoon, S. (2019). Pre-training of deep bidirectional protein sequence representations with structural information.

Moineau, S. (2013). Bacteriophage. In Maloy, S. and Hughes, K., editors, *Brenner's Encyclopedia of Genetics (Second Edition)*, pages 280–283. Academic Press.

Nakashima, T., Ishiguro, N., Yamaguchi, M., Yamauchi, A., Shima, Y., Nozaki, C., Urabe, I., and Yomo, T. (2000). Construction and characterization of phage libraries displaying artificial proteins with random sequences. *Journal of Bioscience and Bioengineering*, 90(3):253–259.

Oloketuyi, S., Bernedo, R., Christmann, A., Borkowska, J., Cazzaniga, G., Schuchmann, H. W., Niedziółka-Jönsson, J., Szot-Karpińska, K., Kolmar, H., and de Marco, A. (2021). Native llama nanobody library panning performed by phage and yeast display provides binders suitable for c-reactive protein detection. *Biosensors*, 11(12).

Olsen, W. L., Staudenbauer, W. L., and Hofschneider, P. H. (1972). Replication of bacteriophage m13: Specificity of the *Escherichia coli* dna b function for replication of double-stranded m13 dna. *Proceedings of the National Academy of Sciences*, 69(9):2570–2573.

Palacios-Rodríguez, Y., Gazarian, T., Huerta, L., and Gazarian, K. (2011). Constrained peptide models from phage display libraries highlighting the cognate epitope-specific potential of the anti-hiv-1 mab 2f5. *Immunology Letters*, 136(1):80–89.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pleiko, K., Põšnograjeva, K., Haugas, M., Paiste, P., Tobi, A., Kurm, K., Riekstina, U., and Teesalu, T. (2020). In vivo phage display: identification of organ-specific peptides using deep sequencing and differential profiling across tissues. *bioRxiv*.

Scikit-Bio (2020). Scikit-bio a bioinformatics library for data scientists, students, and developers. `http://scikit-bio.org`. Online available; accessed at 1/05/2022.

Serrano-Solís, V. and José, M. V. (2013). Flow of information during an evolutionary process: The case of influenza a viruses. *Entropy*, 15(8):3065–3087.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America*, 104(11):4337–4341.

Sidhu, S. S., Weiss, G. A., and Wells, J. A. (2000). High copy display of large proteins on phage for functional selections. *Journal of Molecular Biology*, 296(2):487–495.

Simons, G. F., Konings, R. N., and Schoenmakers, J. G. (1981). Genes vi, vii, and ix of phage m13 code for minor capsid proteins of the virion. *Proceedings of the National Academy of Sciences*, 78(7):4194–4198.

Smith, G. P. (1985). Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317.

Stephen, C. W., Helminen, P., and Lane, D. P. (1995). Characterisation of epitopes on human p53 using phage-displayed peptide libraries: Insights into antibody-peptide interactions. *Journal of Molecular Biology*, 248(1):58–78.

Surge Biswas, M. B. (2019). Unirep protein sequence representations. `https://github.com/churchlab/UniRep`. Online available; accessed at 23/08/2021.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288.

Szabo, F. E. (2015). M. In Szabo, F. E., editor, *The Linear Algebra Survival Guide*, pages 219–233. Academic Press, Boston.

Tarnovitski, N., Matthews, L. J., Sui, J., Gershoni, J. M., and Marasco, W. A. (2006). Mapping a neutralizing epitope on the sars coronavirus spike protein: Computational prediction based on affinity-selected peptides. *Journal of Molecular Biology*, 359(1):190–201.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

White, H. E. and Orlova, E. V. (2019). Bacteriophages: Their structural organisation and function. In Savva, R., editor, *Bacteriophages*, chapter 2. IntechOpen, Rijeka.

Willats, W. G. (2002). Phage display: practicalities and prospects. *Plant Molecular Biology*, 50(6):837–854.

Wu, L., Wen, C., Qin, Y., Yin, H., Tu, Q., Van Nostrand, J. D., Yuan, T., Yuan, M., Deng, Y., and Zhou, J. (2015). Phasing amplicon sequencing on illumina miseq for robust environmental microbial community analysis. *BMC Microbiology*, 15(1):125.

Yang, F., Liu, L., Neuenschwander, P. F., Idell, S., Vankayalapati, R., Jain, K. G., Du, K., Ji, H., and Yi, G. (2022). Phage display-derived peptide for the specific binding of sars-cov-2. *ACS Omega*, 7(4):3203–3211.

Zhong, Y., Cai, J., Zhang, C., Xing, X., Qin, E., He, J., Mao, P., Cheng, J., Liu, K., Xu, D., and Song, H. (2011). Mimotopes selected with neutralizing antibodies against multiple subtypes of influenza a. *Virology journal*, 8:542–542.

APPENDIX

I. Tables

Table 15. PCR primer sequences. Base pairs in capital letters correspond to Illumina adapters, spacers are in italic, target region specific sequences are in lowercase.

Primer name	Primer sequence (5'-3')
Fw1-pair1st	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGattcctttagtggtacctttctattctc
Rw1-pair1st	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG*cgt*caactttcaacagtttcggccg
Fw1-pair2nd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*g*attcctttagtggtacctttctattctc
Rw1-pair2nd	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG*gt*caactttcaacagtttcggccg
Fw1-pair3rd	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*tg*attcctttagtggtacctttctattctc
Rw1-pair3rd	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG*t*caactttcaacagtttcggccg
Fw1-pair4th	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*ctg*attcctttagtggtacctttctattctc
Rw1-pair4th	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcaactttcaacagtttcggccg

Table 16. Phage amplified eluate titering results. BSA concentration in rounds of panning remained constant at 5mg/mL.

	mAb conc. (ug/mL)	Tween-20 (%)	CR3022 (pfu/ul)	H1N1 (pfu/ul)	pAB240 (pfu/ul)	BSA (pfu/ul)
Panning round 1	100	0.1	3.25e+10	4.45e+10	4.40e+10	3.45e+10
Panning round 2	66	0.3	1.60e+10	2.85e+10	4.80e+10	5.50e+10
Panning round 3	33	0.5	7.40e+10	3.45e+9	1.08e+10	4.40e+9

Table 17. Amount of sequences remaining after each filtering step. Control and Naive samples were not used in last two steps since they act as reference. It should be noted that last two pre-processings steps were done on peptide sequences.

Sample name	Stitched reads	Matched fragments	36bp fragments	Without STOP codon	Matching NNK	Quality score	Filtering control sequences	Filtering cross-targets
Naive non-amp.	3759060	2950321	2938795	2780438	2736668	2335002	-	-
Naive amp.	3558467	2510499	2501584	2410480	2370366	2010640	-	-
CR3022 Pan. 1	392575	306321	305381	298546	295668	275881	139490	137158
CR3022 Pan. 2	276894	252016	251792	249025	248219	231804	33250	17829
CR3022 Pan. 3	287520	268291	268116	264573	263953	245144	33916	13418
H1N1 Pan. 1	375999	287520	285389	277057	272909	253230	242666	214906
H1N1 Pan. 2	343098	265488	264945	277057	272909	253230	199786	108081
H1N1 Pan. 3	375673	202973	202624	200792	167392	154800	110218	27943
PAB240 Pan. 1	366296	292775	291646	284909	281074	259827	256046	245353
PAB240 Pan. 2	382686	285436	284478	279960	276828	255756	203291	186799
PAB240 Pan. 3	409210	225617	225051	223116	2211527	205194	251212	227520
Control (BSA) 1	419081	310665	309347	300753	297077	274983	-	-
Control (BSA) 2	320610	218716	218443	217840	217035	200146	-	-
Control (BSA) 3	434675	226218	226058	225930	225450	205317	-	-

Table 18. Detailed description of phage libraries across mimosets in Dataset A.

Target	Mimoset	Randomness	Phage Library name	Library type	Peptide length	Diversity	Article ref., PMID
1	1	Semi-random	LX8 and X15CX	Linear	14	-	16940148
1	2	Completely random	X15 and X21	Linear	15 or 21	-	16940148
1	3	Completely random	CX12C	Circular	12	-	16940148
1	4	Completely random	f3-15mer	Linear	12	$2.5x10^8$	9430247
1	5	Completely random	X21	Linear	22	$6.5x10^8$	9430247
1	6	Semi-random	XCX(3)SDLX(3)CI	Circular	10	-	11413337
1	7	Semi-random	X(7)SDLX(3)CI	Circular	14	-	11413337
2	1	Completely random	f3-15mer and f88-15mer	Linear	15	-	16630634
2	2	Semi-random	f88-Cys1	Circular	13	$2.8x10^9$	16630634
3	1	Completely random	J404	Linear	9	-	17675514
4	1	Completely random	X20	Linear	16	$1x10^9$	8798975
4	2	Completely random	X20	Linear	20	$5x10^9$	19653209
4	3	Semi-random	X16	Linear	7	-	22870226
4	4	Semi-random	X16	Linear	7	-	22870226
4	5	Semi-random	X16	Linear	7	-	22870226
5	1	Completely random	Ph.D.-12	Linear	7	$1x10^9$	21237206
5	2	Completely random	Ph.D.-C7C	Circular	12	$1x10^9$	21237206
6	1	Completely random	X10	Linear	10	-	18855146
6	2	Completely random	X10	Linear	10	-	18855146

Table 19. Full target names according to the number used in the work.

Target Number	Dataset A target name	Dataset B target name
1	Anti-gp120 monoclonal antibody b12	Anti-HIV-1 gp41 MPER monoclonal antibody 2F5
2	Anti-spike glycoprotein monoclonal antibody 80R	Anti-MSP1a monoclonal antibody 15D2
3	Anti-FPR monoclonal antibody NFPR1	Anti-NTX monoclonal antibody BNTX18
4	Anti-gp120 monoclonal antibody GV4H3	Anti-HSP90 monoclonal antibody AC88
5	Anti-HIV-1 gp41 MPER monoclonal antibody 2F5	Anti-coagulation factor VIII monoclonal antibody 2-76
6	Anti-p53 monoclonal antibody	Anti-glycoprotein E2 monoclonal antibody 3/11

Table 20. Dataset A hierarchical clustering detailed results. Peptide distribution across different clusters based on their mimoset origin. As example, "Target.Mimoset" value 1.3 should be read as mimoset 3 belonging to target 1.

Cluster	1	2	3	4	5	6
Target.Mimoset						
1.1				2		
1.2				31		1
1.3				3		16
1.4				3		
1.5				2		
1.6			6			2
1.7						10
2.1				17		1
2.2				36	3	2
3.1		86				4
4.1				8	1	
4.2				6		
4.3				13	1	6
4.4				8	9	
4.5				4	3	1
5.1					42	2
5.2	27					
6.1			23	5		
6.2			30	1		3

Table 21. Detailed description of phage libraries across mimosets in Dataset B.

Target	Mimoset	Randomness	Phage Library name	Library type	Peptide length	Diversity	Article ref., PMID
1	1	Completely random	Ph.D-12	Linear	12	10^9	21237206
2	1	Completely random	Ph.D-12	Linear	12	10^9	22427942
3	1	Completely random	Ph.D-12	Linear	12	10^9	11275260
4	1	Completely random	Ph.D-12	Linear	12	10^9	19741295
4	2	Completely random	Ph.D-12	Linear	12	10^9	19741295
5	1	Completely random	Ph.D-12	Linear	12	10^9	25520269
6	1	Completely random	Ph.D-12	Linear	12	10^9	16496330
6	2	Completely random	Ph.D-12	Linear	12	10^9	16496330

Table 22. Peptide distribution across different clusters based on their mimoset origin in Dataset B. As example, "Target.Mimoset" value 1.3 should be read as mimoset 3 belonging to target 1.

Cluster	1	2	3	4	5	6	7
Target.Mimoset							
1.1	45						
2.1			2		2	34	1
3.1			8	13	11	1	
4.1							15
4.2							12
5.1		23	1	2	1		
6.1			1	3	15		
6.2				1	4		

II. Figures

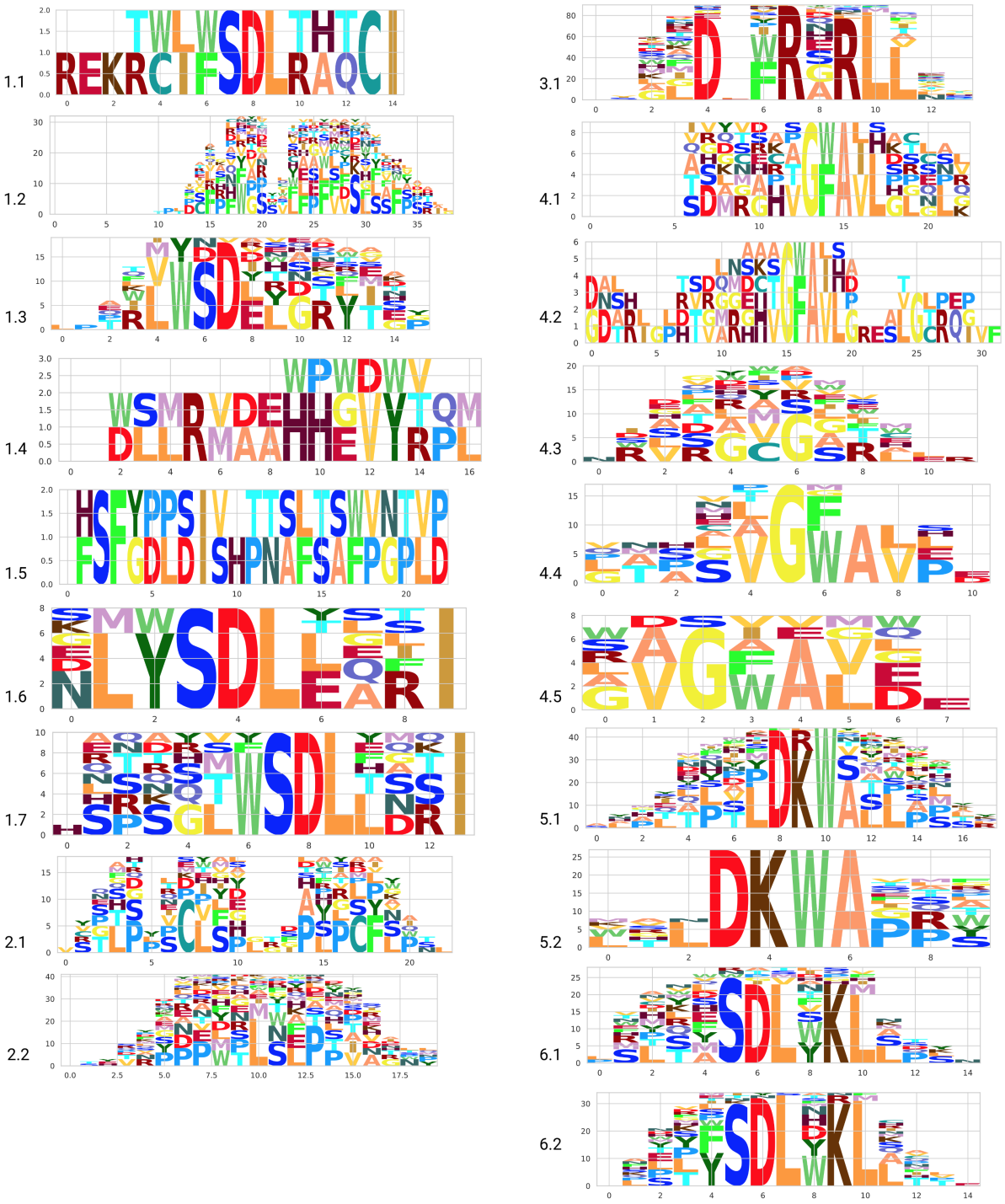

Figure 18. Multiple sequence alignment logos for each of the mimosets in the Dataset A. Horizontal axis show alignment position, vertical axis show number of occurrences of the aligned residue.

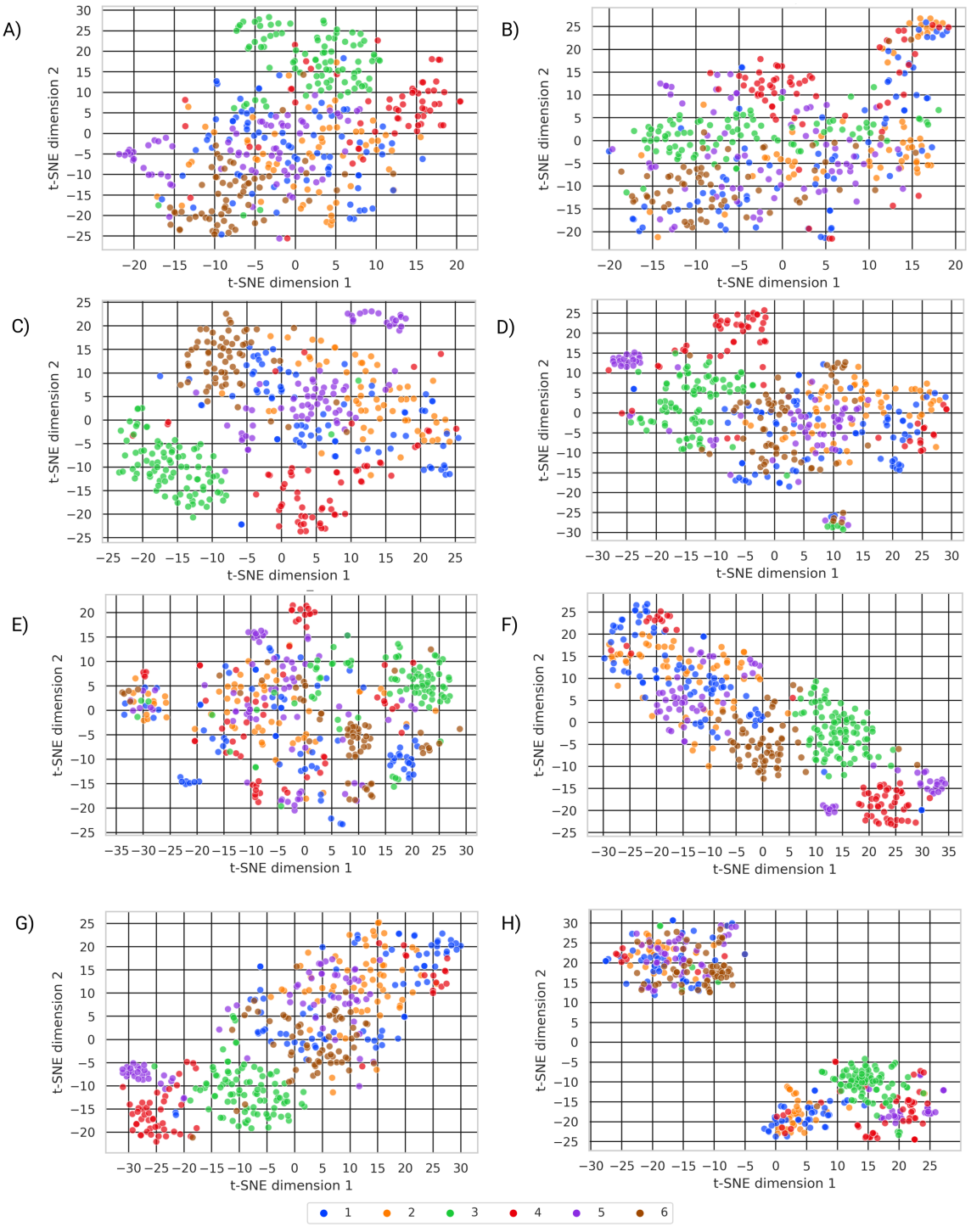

Figure 19. Visualization of t-SNE dimensionality reduced peptide representations generated by various methods on Dataset A. A) Peptide representations generated by BERT, B) PLUS, C) SeqVec, D) UniRep h_avg, E) UniRep h_final, F) RGN_f1 layer, G)RGN_r1 layer, H) RGN_r2 layer.

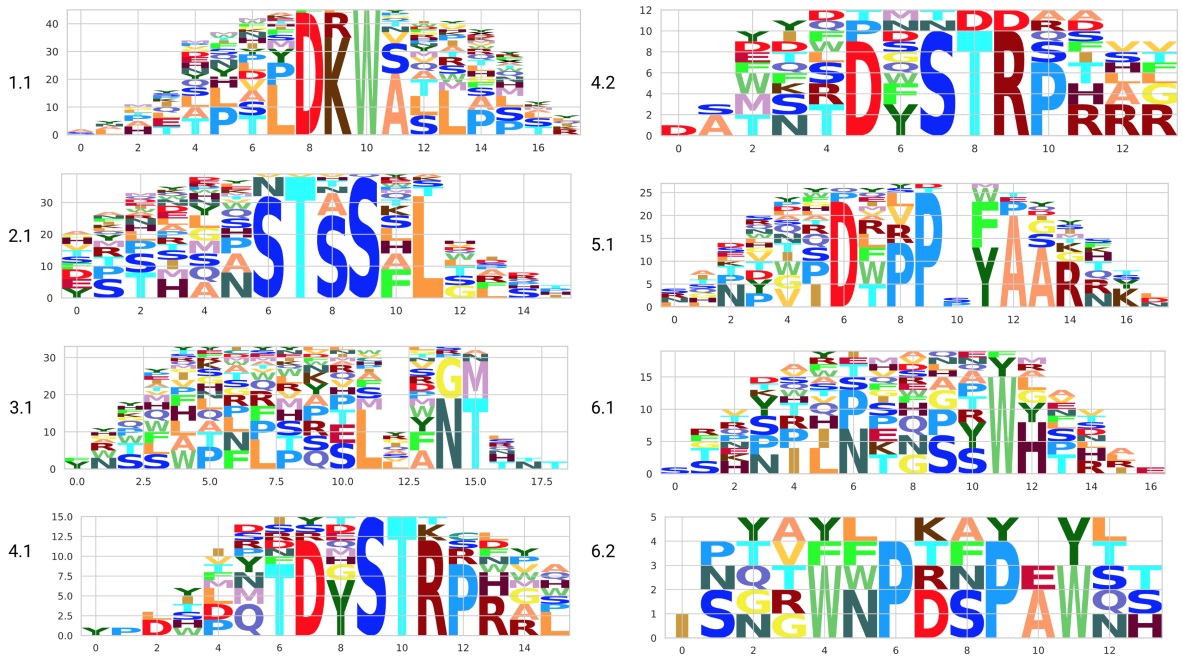

Figure 20. Multiple sequence alignment logos for each of the mimosets in the Dataset B. Horizontal axis show alignment position, vertical axis show number of occurrences of the aligned residue.

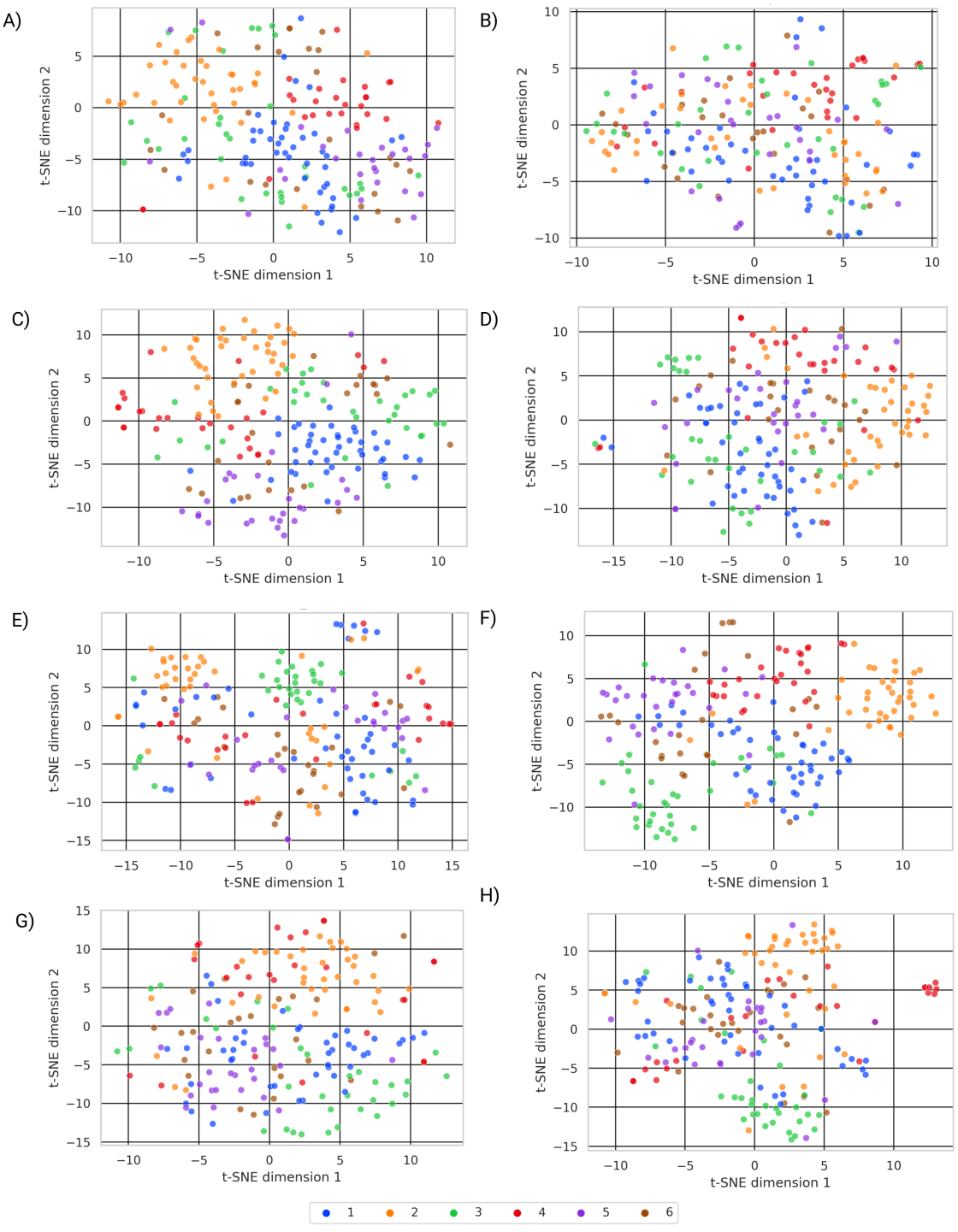

Figure 21. Visualization of t-SNE dimensionality reduced peptide representations generated by various methods on Dataset B. A) Peptide representations generated by BERT, B) PLUS, C) SeqVec, D) UniRep h_avg, E) UniRep h_final, F) RGN_f1 layer, G)RGN_r1 layer, H) RGN_r2 layer.

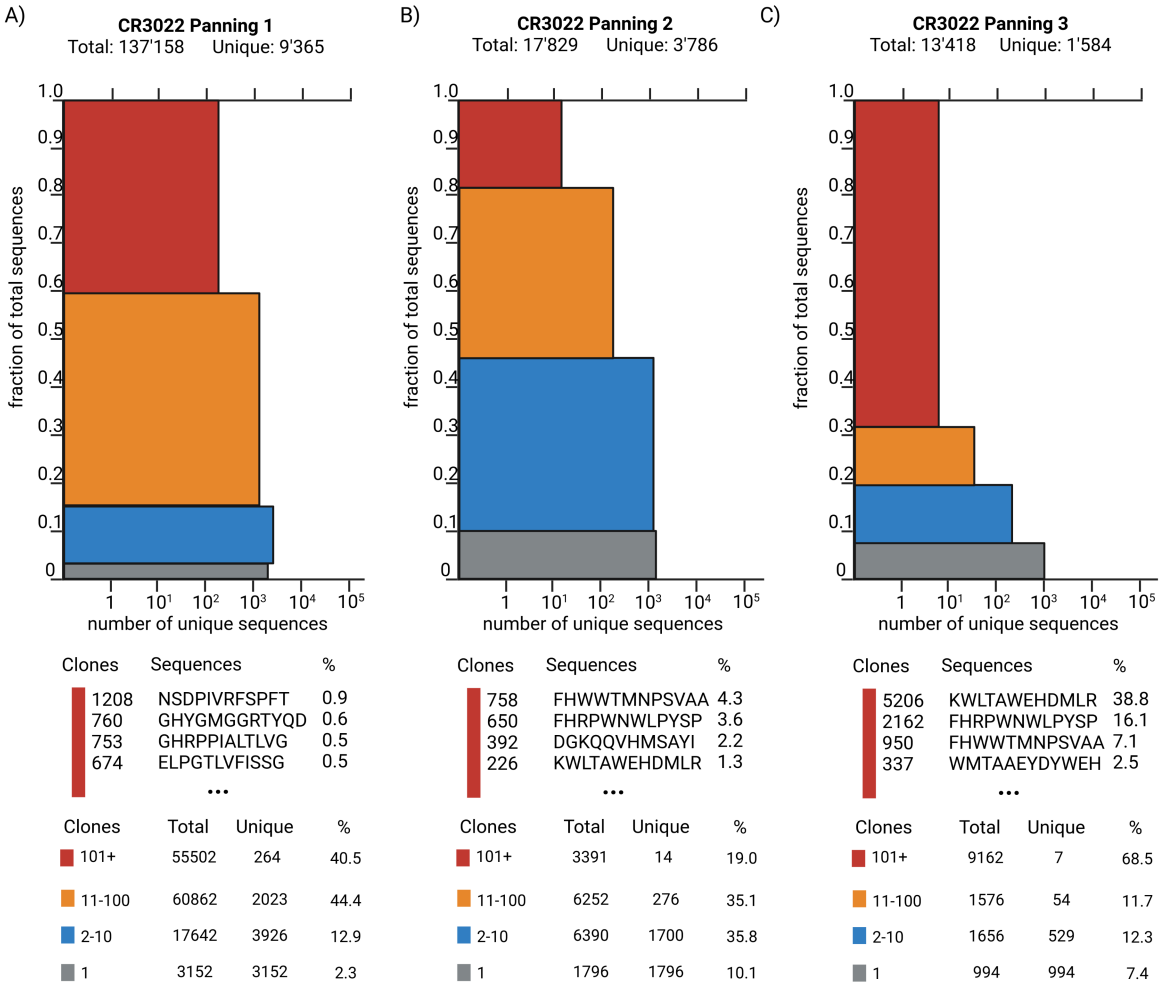

Figure 22. Stacked bar plot of pre-processed sequencing data from CR3022 mAb sample. Y-axis shows the fraction from total sequences that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total sequence pool.

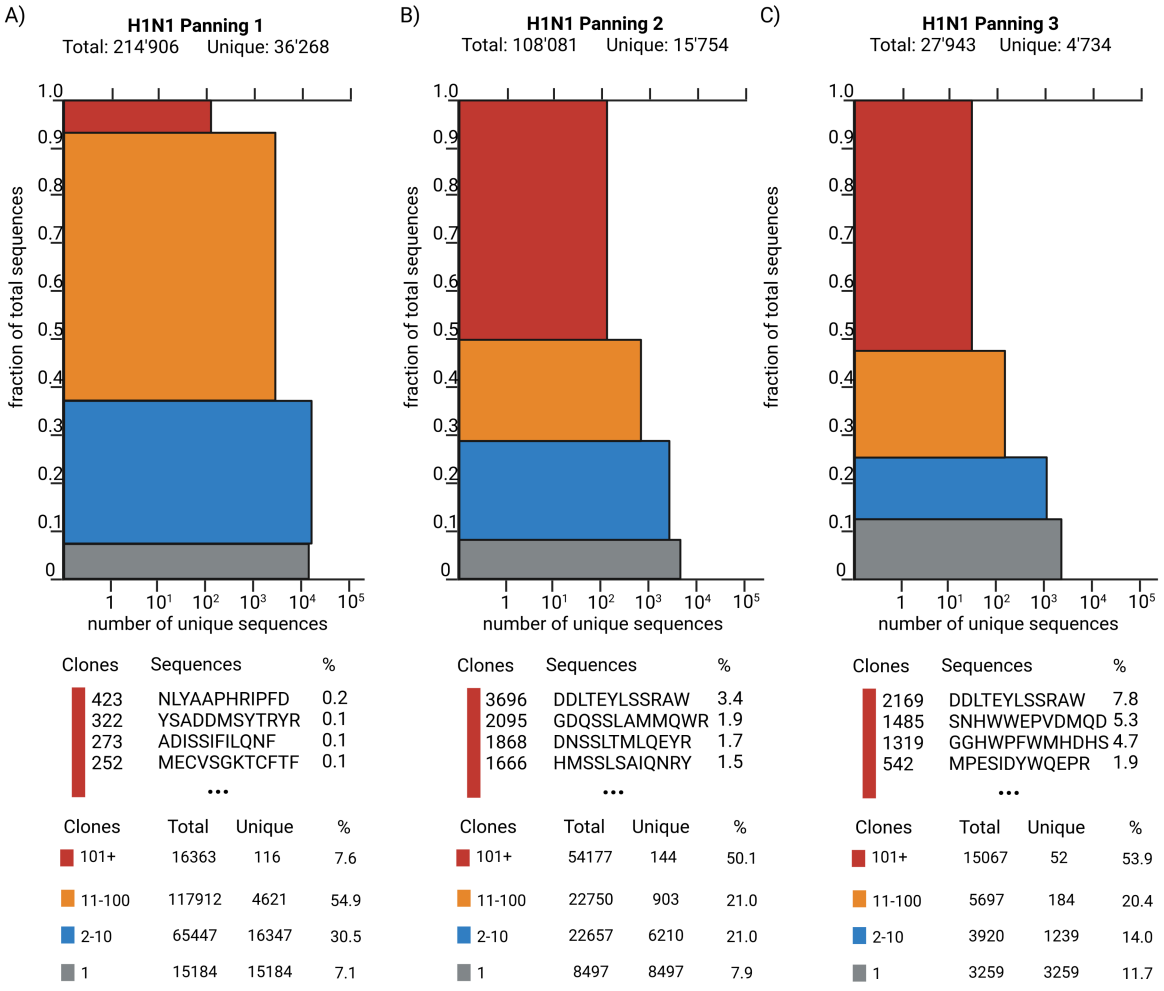

Figure 23. Stacked bar plot of pre-processed sequencing data from H1N1 mAb sample. Y-axis shows the fraction from total sequences that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total sequence pool.

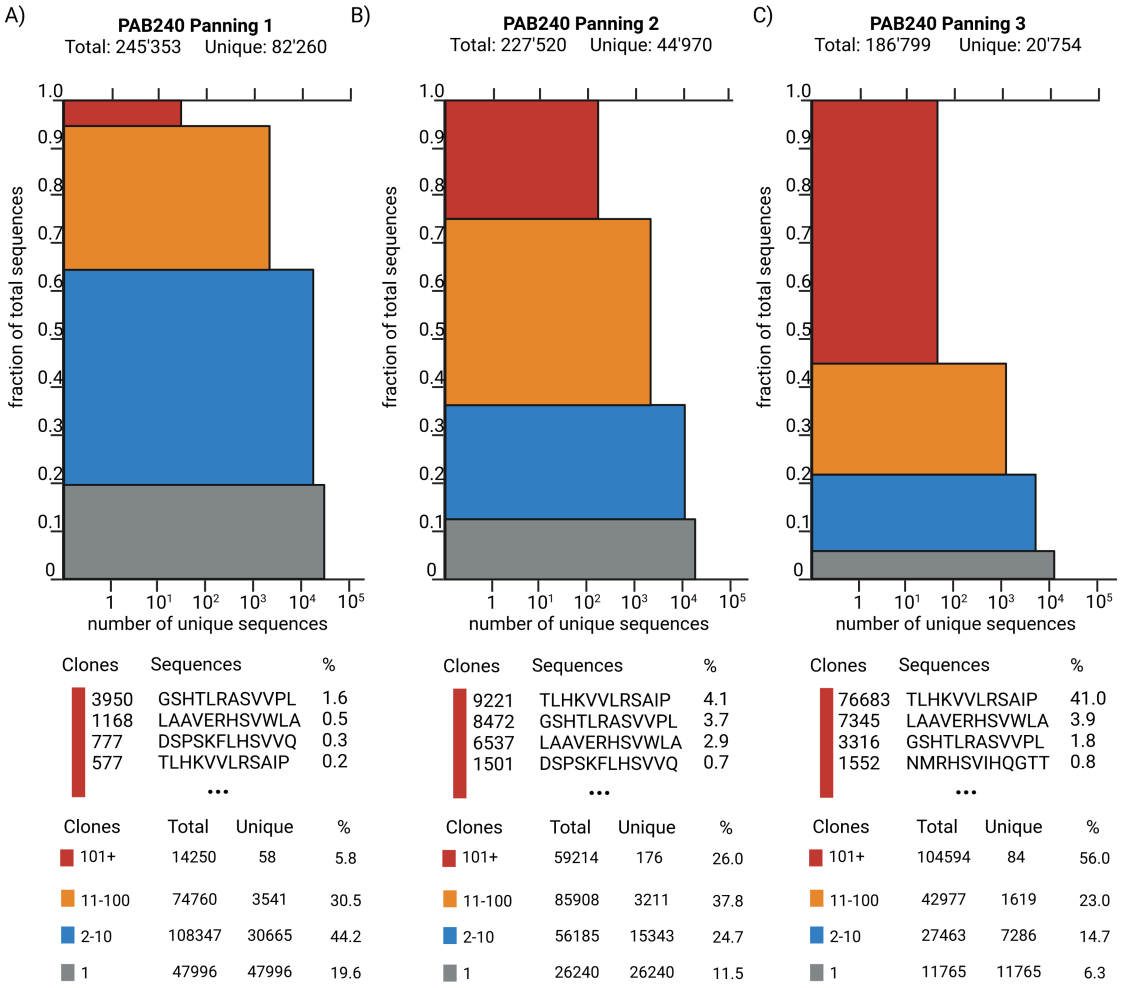

Figure 24. Stacked bar plot of pre-processed sequencing data from PAB240 mAb sample. Y-axis shows the fraction from total sequences that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total sequence pool.

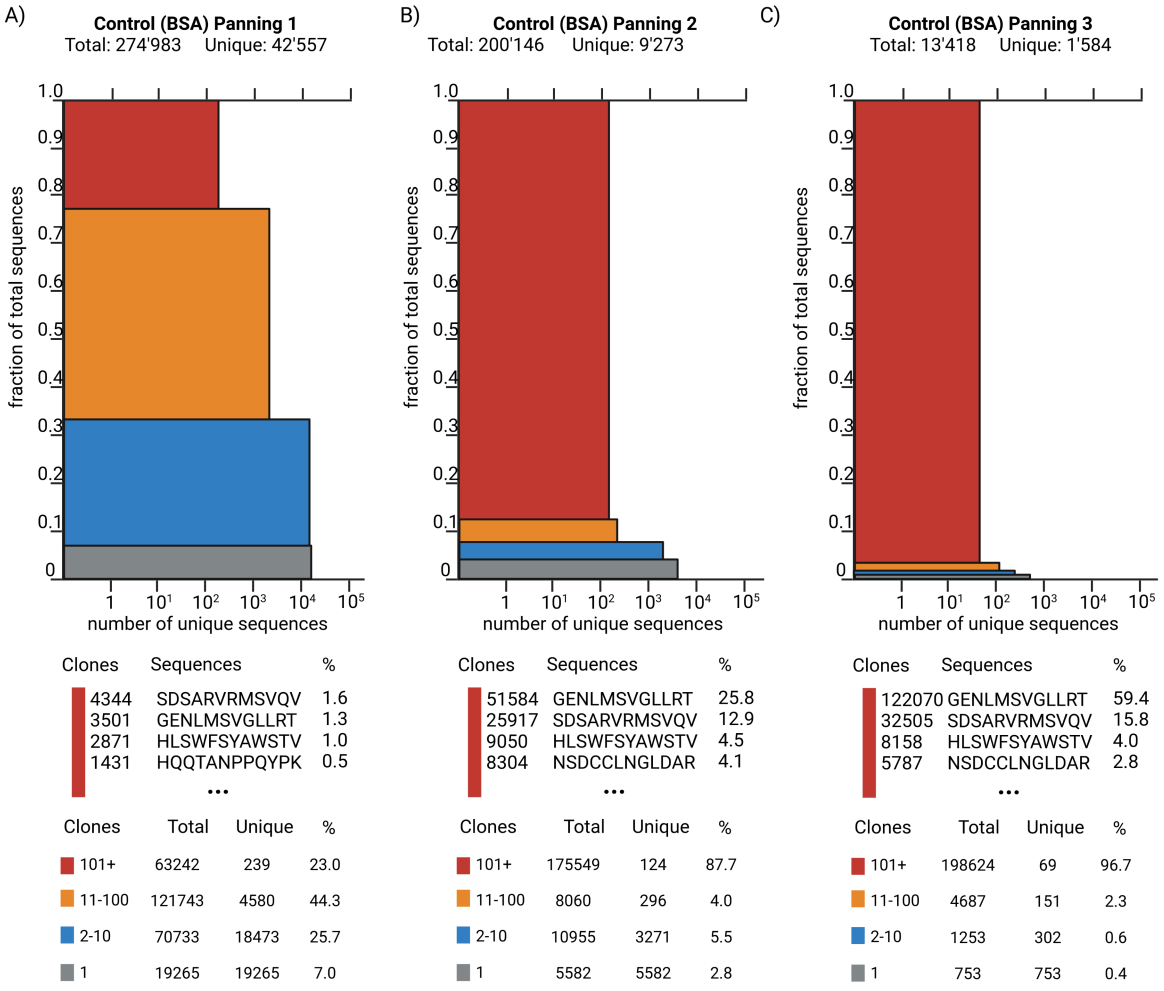

Figure 25. Stacked bar plot of pre-processed sequencing data from control (BSA) sample. Y-axis shows the fraction from total sequences that is occupied by the clones while X-axis shows the amount of unique sequences among the clones. Table highlights most abundant sequences and their fractions from total sequence pool.

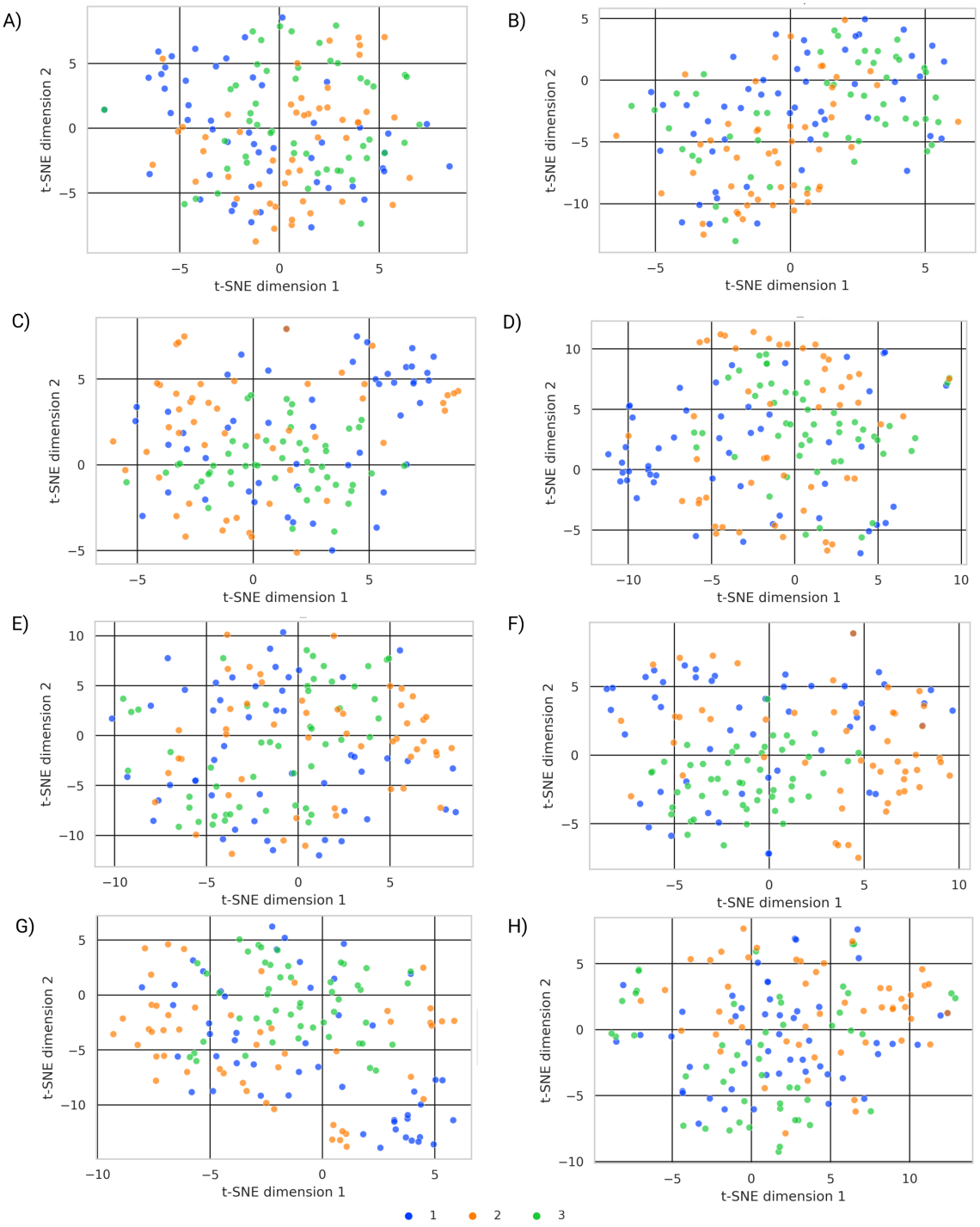

Figure 26. Visualization of t-SNE dimensionality reduced peptide representations generated by various methods on obtained experimental data. A) Peptide representations generated by BERT, B) PLUS, C) SeqVec, D) UniRep h_avg, E) UniRep h_final, F) RGN_f1 layer, G)RGN_r1 layer, H) RGN_r2 layer. Legend; 1 - CR3022, 2 - H1N1, 3- PAB240.

III. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Klavs Jermakovs**,

(author's name)

1. 1. grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

   ***In vitro* and *in silico* epitope-paratope mapping**,

   (title of thesis)

   supervised by Tuomas Knowles, Erik Abner and Alekszej Morgunov.

   (supervisor's name)

2. I grant the University of Tartu the permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from **27/05/2022** until the expiry of the term of copyright,

3. I am aware that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Klavs Jermakovs

27/05/2022