Incorporating machine learning into a VHH candidate discovery pipeline

Ben Wagstaffe I, Anna Boczkaj II, Marion Cubitt II, Bill Eldridge II, Anna Hine III, Yordan Raykov I.

I School Of Mathematical Sciences, The University Of Nottingham, University Park, Nottingham NG7 2RD

II Isogenica Ltd, The Mansion, Chesterford Research Park, Cambridge, UK, CB10 1XL

III College and Health and Life Sciences, Aston University. Birmingham. B4 7ET

This project is supported by a KTP grant from Innovate UK.

INTRODUCTION

VHHs are small-format single domain antibodies with a multitude of pharmaceutical applications thanks to their small size, specificity and affinity. This KTP (Knowledge Transfer Partnership) project combines the knowledge and experience from two universities (Aston University and their DNA library build methodology and University of Nottingham's Machine Learning (ML) capabilities) with the antibody screening and engineering expertise of Isogenica Ltd.

The project aims to develop and integrate an ML-assisted pipeline for VHH discovery and optimization that adapts quickly to new molecular targets, without relying on access to structural information about the antigen or antibody. Instead, the pipeline efficiently queries predefined VHH libraries and predicts residue-specific and antibody-specific scores. These scores will allow antibody ranking based on binding affinity and

WHAT ARE VHHS?

- Small single domain antibodies (approx. 12-15 kDa) originally derived from camelids, with intrinsically high solubility[2] and low immunogenicity[3].
- Wide range of biopharmaceutical applications, including multi-specific antibody therapeutics, ADCs, immune engagers, CARs, radiopharmaceuticals and more.
- VHH variable complementarity-determining region (CDR) domains are responsible for their binding attributes (Figure 1). The CDR3 loop is typically longer than conventional IgG antibodies[4],enabling access to cryptic epitopes.

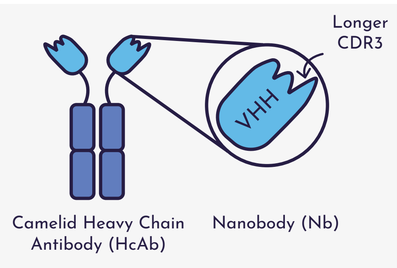
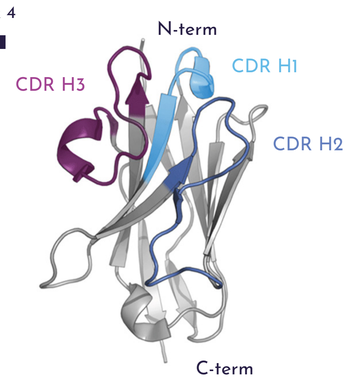


Fig. 1: *Schematic of a typical VHH antibody structure and sequence, highlighting the variable CDR loops and regions of sequence diversity[1].*



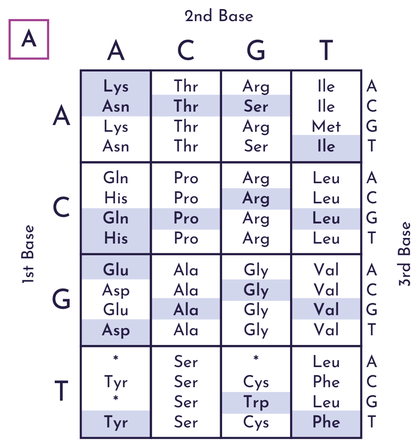
1. GENERATING AND SCREENING VHH LIBRARIES FOR ML-ASSISTED DISCOVERY

VHH sequences from immune libraries have a significant number of variations spanning both the conserved framework and CDR domains. Synthetic libraires encoding a fixed framework with reduced CDR variability allows a more controlled approach to applying ML tools for VHH discovery and engineering.

Isogenica's existing synthetic VHH libraries possess a huge amount of CDR diversity. These vast libraries(~10^{12}) are greater than the cloneable size for phage display screening, making the application of ML in antibody discovery more challenging.

We are generating focused fully synthetic libraries designed using the wealth of knowledge at Isogenica gathered through many years of VHH discovery. Non-degenerate saturation mutagenesis will enable efficient use of sequence space, creating VHH libraries of a manageable size that can be fully cloned for phage display while still achieving appropriate CDR diversity (Figure 2).

Screening these focused synthetic libraries and subsequent NGS sequencing of positive hits will allow greater certainty in the dataset – key for successful ML training – as all VHHs in the library will be represented in either the positive or negative hits from the screen (Figure 3). We can then use ML approaches to cluster positive hits and home in on the most promising leads.



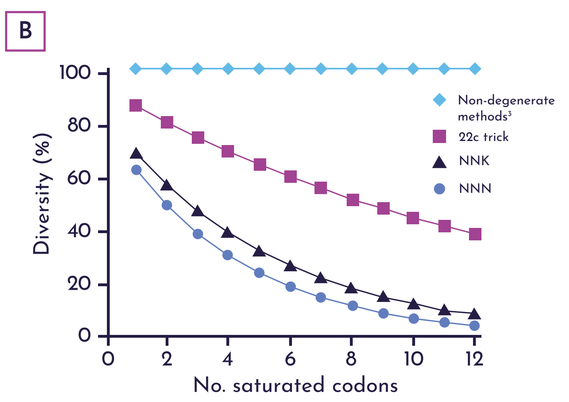


Fig. 2: *A: The concept of non-degenerate saturation where one amino acid is encoded by one codon. B: A graph showing the significant impact on diversity of using degenerate saturation mutagenesis techniques[5].*

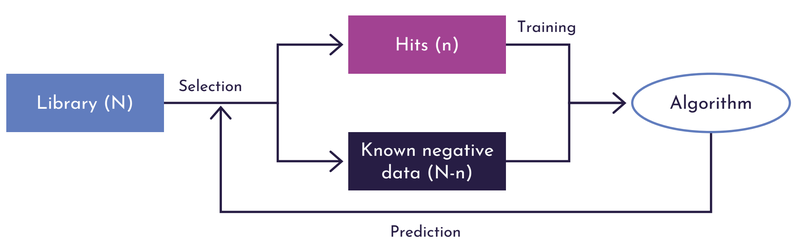


Fig. 3: *Working with a cloneable sequence space allows inference of negative data as well as positive data. This can enhance iterative training of a predictive algorithm.*

2. ML APPROACHES FOR VHH OPTIMISATION

Once positive hits have been identified and sequenced, additional protein engineering may be necessary to further optimize their binding and characteristics.

ML approaches could improve the speed and efficiency of antibody optimization by predicting how best to modify an antibody to bind its target.

However, structural information about antibodies and their antigen binding targets is not easily accessible. And without large libraries of structural data, we cannot train conventional data-hungry neural architectures.

Instead, we are developing a few-shot learning knowledge graph that predicts position-specific binding affinity score and overall antibody naturalness score using only limited labelled data and larger amounts of meta data derived from existing atlases of structural information such as the Protein Data Bank (Figure 4)[6].

By repurposing established interaction motifs estimated from known protein structures, we extract pocket definitions of frequent interactions which are used as a semi-supervised prototypical knowledge in a graph model that regresses binding affinity scores for new antibodies (Figure 5).

Edge weights and additional constraints incorporate domain knowledge and learn to embed unique information from the focused libraires.

VHH antibodies will be clustered to prioritize screening of functionally different antibodies against a target and prototype binding affinities will be used to predict a score directly from sequence data about a target and a selected VHH.

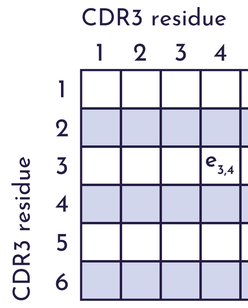


Fig. 4: Example CDR3 matrix

Each $e_{i,j}$ denotes a vector with edge features, such as structural information (pairwise atom distances and dihedral angles), property information (solubility, co-occurrence motifs) and VHH library design biases.

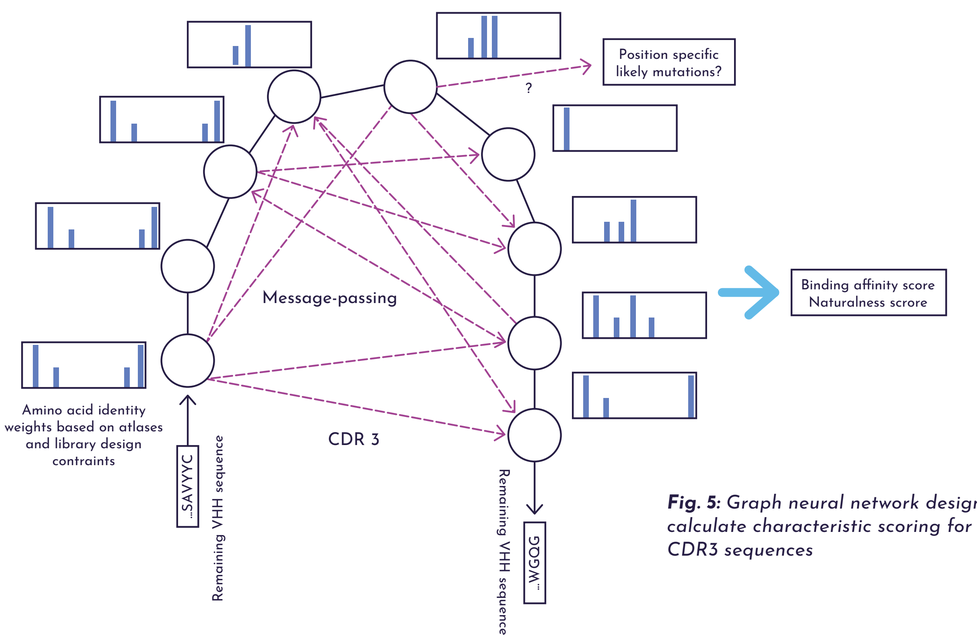


Fig. 5: *Graph neural network design to calculate characteristic scoring for variable CDR3 sequences*

SUMMARY

- **Non-degenerate saturation mutagenesis is essential for efficient use of sequence space, generating fully cloneable VHH libraries with high CDR diversity for phage display screening**
- **Producing fully screenable libraries allows greater certainty in negative as well as positive data for ML training to identify the most promising leads**
- **Atlases of frequent protein interactions can be used to score residue identities in CDRs**
- **Contrast between pairwise interactions and design frequency of different amino acids in a VHH library can be fused together in a graph model which can address residue specific and network specific queries to predict optimal antibody-target interactions**
- **ML can support efficient identification and optimisation of VHH antibodies for multiple therapeutic applications**

References: 1: Eliseev IE et al. (2018) F1000Research 7 (57) 2: Conrath K et al. (2005) J Mol Biol. 350(1): 112-125 3: Ackaert C et al. (2021) Front. Immunol. 12: 632687 4: Desmyter A et al. (1996) Nat Struct Biol. 3(9): 803-811 5: Ashraf M et al. (2013) Biochem Soc Trans. 41(5): 1189-94 6: Kynast, J.P. et al. (2022) ATLIGATOR. bioRxiv.

ISOGENICA HQ: THE MANSION HOUSE

Based at Chesterford Research Park,
Cambridge CB10 1XL, UK

T +44 1799 533 680
E bd@isogenica.com

To learn more about Isogenica Ltd, visit our website