AbNatiV: VQ-VAE-based assessment of antibody and nanobody nativeness for engineering, selection, and computational design

Aubin Ramon[1], Alessio Saturnino[1,2], Kieran Didi[1,3,a], Matthew Greenig[1], Pietro Sormanni[1*]

[1] Centre for Misfolding Diseases, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield road, CB2 1EW Cambridge, UK

[2] Department of Biology and Biotechnology Lazzaro Spallanzani, University of Pavia, Via Adolfo Ferrata 9, 27100 Pavia, Italy

[3] Faculty of Biosciences, Heidelberg University, Im Neuenheimer Feld 234, 69120 Heidelberg, Germany

[a] Current address: Faculty of Mathematics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK

* Corresponding Author: ps589@cam.ac.uk

Keywords: Antibody engineering, antibody design, deep learning, humanization, humanness, immunogenicity, machine learning, nanobody, vector-quantized variational autoencoder.

Abstract

Monoclonal antibodies have emerged as a key class of therapeutics, and nanobodies are rapidly increasing in popularity following the approval of the first nanobody drug in 2019, yet the therapeutic development of these biologics remains a challenge. Computational design is a promising technology to accelerate antibody discovery and optimisation and to address some critical limitations. However, despite the availability of established in vitro directed evolution technologies that are relatively fast and cheap to deploy, the gold standard for generating therapeutic antibodies remains discovery from animal immunization or patients. Immune-system derived antibodies tend to have favourable properties in vivo, including long half-life, low cross reactivity with self-antigens, and low toxicity, which raises the question of whether computational design will ever be able to generate such antibodies. Here, we present AbNatiV, a deep-learning tool for assessing the nativeness of antibodies and nanobodies, i.e., their likelihood of belonging to the distribution of immune-system derived human antibodies or camelid nanobodies. AbNatiV is trained on curated datasets of human antibodies or camelid nanobodies and can accurately predict the nativeness of a given sequence. We show that AbNatiV can be used to design antibodies and nanobodies that are indistinguishable from immune-system derived ones, to facilitate humanization, and to predict the likelihood of immunogenicity. AbNatiV is a multi-purpose tool that will be valuable for developing antibodies and nanobodies with improved efficacy and safety profiles. We make it readily accessible to users through both downloadable software and as a webserver.

1 Introduction

Antibodies are a highly valuable class of biomolecules with the remarkable ability to bind to molecular targets selectively and tightly. For this reason, they find key applications in biological research [1] and medicine, where they are widely employed both as diagnostic [2] and as therapeutic agents [3]. Nanobodies (Nbs) are single-domain antibodies (VHH) naturally expressed in camelids [4]. They have grown in popularity due to their unique structural characteristics, which include small size, good stability and solubility, long third complementarity determining region (CDR3) that can bind to poorly accessible epitopes, and affinity and specificity at par to those of full-length antibodies [5]. Furthermore, their employment as therapeutics has gained increased momentum since the approval of Caplacizumab in 2019, the first nanobody therapeutic [6].

Established approaches to discover new antibodies or nanobodies for a target of interest can broadly be classified as first-generation in vivo approaches, for instance relying on animal immunisation [7], and second-generation in vitro techniques, relying on laboratory library construction and screening [8]. More recently, a third generation of approaches based on computational design has started to emerge [9]. Starting from the mid 90s, in vitro methods like phage display from naïve or synthetic libraries showed promise to replace animal immunisation or other in vivo techniques to isolate novel antibodies. In vitro selection is faster and cheaper than in vivo counterparts, has no ethical implications, and enables a better control over antigen presentation [10,11]. However, an increasing number of pharmaceutical and biotech companies prefers to obtain new antibodies by immunising transgenic animals with a humanised immune system [12,13] or by isolating them directly from patients [14,15], in spite of the added costs and complexity. The reason for this choice is that, compared with in vitro directed evolution, antibody selection carried out by immune systems usually yields antibodies with better developability potential and especially in vivo properties, including long half-life, low immunogenicity, no toxicity, and low cross-reactivity against self-antigens [16,17]. Up to now, the vast majority of therapeutic antibodies continues to come from animal immunization [18]. This consideration thus raises the question of whether a computational design strategy will ever rise to meet the challenge of generating antibodies with such properties.

Computational antibody design is still in its infancy. Yet, important advances have been made in the design of antibodies targeting predetermined epitopes of interest [19–23], which remains extremely laborious with laboratory-based approaches, and in the prediction and design of biophysical properties that underpin developability [24]. Overall, computational design promises a cheaper and faster route for the discovery and optimisation of antibodies, while in principle affording a much better control over other key biophysical properties such as stability and solubility than in vivo and in vitro techniques [9].

Notwithstanding these advances, the computational prediction of in vivo properties remains hugely problematic. These properties, which include long half-life, low immunogenicity, and no toxicity, are difficult to measure accurately and in good throughput, and their molecular determinants remain poorly understood. This hurdle broadly affects therapeutic antibody development also beyond computational design, and a multitude of in vitro assays, referred to as developability screening assays, have been proposed as proxies for binding specificity or in vivo half-life to de-risk antibody development programmes [24–26]. However, these assays typically correlate poorly with each other, and have only been shown to somewhat correlate with selected in vivo properties in limited specific examples [17,24,27]. While advances have been made in the computational predictions of the outcome of some of these assays [28–30], or even in the number of such assays in which a lead antibody candidate is likely to perform poorly [31,32], it is quite clear that progress is hindered by the absence of robust well-defined experimental measurements of in vivo properties. These challenges are the key reasons behind

the fact that in vivo antibody discovery from immune systems largely remains the gold-standard technology for therapeutic antibody discovery.

In this work, we introduce a novel deep learning method to bypass these challenges, by enabling the computational design of antibody sequences indistinguishable from those obtained from immune systems. We call our method AbNatiV, as it provides an accurate quantification of the likelihood of a given sequence belonging to the distribution of those of native variable domains (Fv) derived from human or camelid immune systems. We define this likelihood antibody nativeness, as it reflects the similarity to native antibodies. Therefore, Fv sequences with high nativeness can be expected to have in vivo properties comparable to those of immune-system-derived antibodies. AbNatiV consists of a vector-quantized variational auto-encoder (VQ-VAE) designed to process aligned Fv sequences and trained with masked unsupervised learning on sequences from curated native immune repertoires. Four different models are trained respectively on the Fv sequences of human heavy chains (VH), kappa light chains (VKappa), lambda light chains (VLamda), and camelid heavy chain single-domains (VHH, that is to say nanobodies).

AbNatiV can assess the degree of humanness and of VHH-nativeness of a given Fv sequence. It provides both an interpretable overall nativeness score and a residue-level nativeness profile. Therefore, AbNatiV can be very useful for computational antibody design, but also to rank Fv sequences of any origin, including from in vitro discovery. Besides, the residue-level nativeness profile can readily guide the engineering of nativeness-improving mutations. We find that AbNatiV learns a representation of natural antibodies that captures high-order relationships between positions to provide a highly sensitive measure of antibody nativeness. For example, it provides a precise assessment of the humanness of antibodies, including in the context of therapeutic development when compared to other methods developed for antibody humanisation. AbNatiV also defines the concept of VHH-nativeness and introduces a novel approach to monitor concurrently in silico the humanness and the VHH-nativeness of a nanobody, which can assist the growing field of nanobody development.

2 Materials and Methods

2.1 Datasets and antibody sequence processing

The source of all antibody sequences used for training and testing is given in **Table S1**, with most sequences coming from the Observed Antibody Space (OAS) [33]. All sequences were aligned, cleaned, and processed beforehand. Non-redundant sequences were aligned using the AHo numbering scheme [34] resulting in aligned sequences of length 149. The alignment was carried out using the ANARCI software [35] followed by a custom python script to check for consistency and fix misalignments. More specifically, we found that in some instances gaps may be opened in unexpected positions (sometimes in framework 1 or framework 2) leading to a misalignment of the subsequent part of the sequence, including the fully conserved cysteines that form the intra-domain disulphide bond (AHo positions 23 and 106). Therefore, a script was run to adjust possible inaccuracies in the alignment of each sequence within the multiple sequence alignment (MSA). This script maximises the identity between the MSA consensus sequence and the sequence under scrutiny calculated at all positions with conservation index greater than 0.9, which include the two fully conserved cysteines. Sequences whose alignment could not be fixed, or that didn't have two cysteines at the conserved positions (because of e.g., sequencing errors) were discarded. Furthermore, Fv sequences with more than one or two missing residues at respectively the N- and the C-terminal were removed. For heavy chains, a Glutamine residue was added at the N-terminus, if missing, and two Serine residues were added at the C-terminus, if missing. For lambda and kappa light chains, respectively a Leucine or a Lysine were added at the C-terminus (AHo

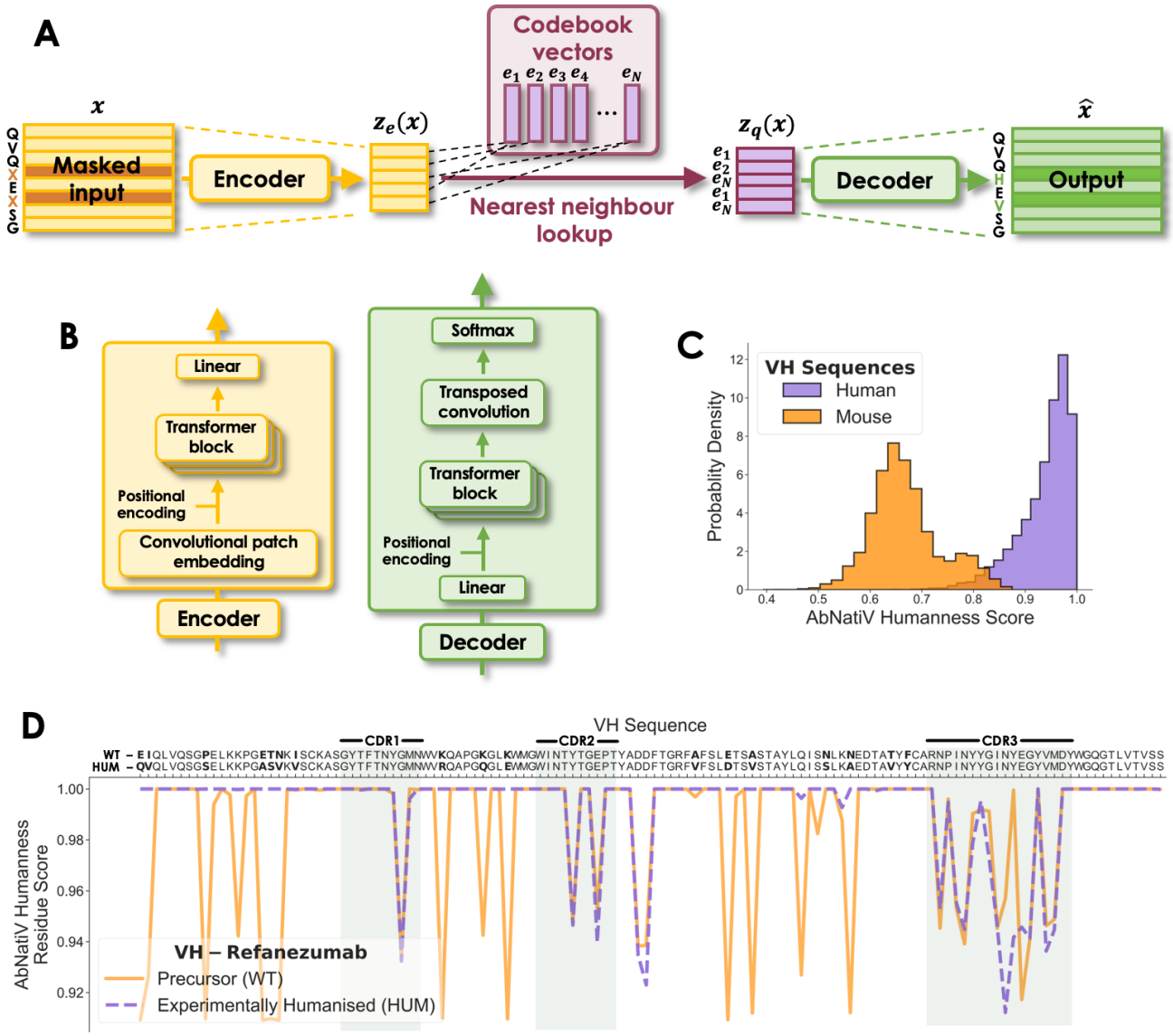

Figure 1. The AbNatiV model. (**A**) Architecture of the VQ-VAE-based AbNatiV model. The one-hot encoded input sequence x is encoded into a compressed representation $z_e(x)$ through an encoder (in yellow). In the latent space (in burgundy), $z_e(x)$ is discretised with a nearest neighbour lookup on a codebook $\{e_k\}_{k=1}^{N}$ of N code-vectors. Each of the components of $z_e(x)$ is substituted with the closest code-vector to generate the discrete embedding $z_q(x)$. Finally, the output $\hat{x}$ is reconstructed through a decoder (in green) from $z_q(x)$. During training, residue masking is applied to the input x by replacing a portion of its residues with a masking vector (in darker shade). (**B**) Architecture of the encoder (in yellow) and decoder (in green) blocks in the AbNatiV model. (**C**) AbNatiV humanness score distributions of the VH Human (Test set, in purple) and Mouse databases (in orange). The ROC-AUC between the two distributions is 0.996. (**D**) AbNAtiV humanness profiles of the VH mouse precursor and of the humanised sequence of the Refanezumab antibody therapeutic (the corresponding VL profile is in **Fig. S2**).

position 148), if missing. After alignment a check for unique sequences was repeated (because for example after completing the C-terminus some duplicated sequences may exist) and any duplicate discarded.

Datasets of processed heavy, lambda, kappa and VHH antibody sequences from various studies from the literature were assembled (**Table S1**) and processed as described above. 2,000,000 sequences from the human heavy, lambda, and kappa databases were used to train three distinct models respectively. 2,127,676 sequences from single-domain antibodies were used to train a fourth model. For each model, 50,000 sequences were additionally kept aside for validation, and 10,000 sequences for testing. Position weight matrices (PWM) and corresponding position-specific scoring matrices (PSSMs) were computed for each human and camelid antibody datasets (**Fig. S1**). From these matrices, additional custom datasets of artificial sequences were generated to be used as controls, named PSSM-generated datasets. These sequences were built by randomly filling each residue position using the underlying residue frequency observed in the PWM (that is the matrix of observed residue frequencies, **Fig. S1**) considering only those amino acids enriched at that position (that is PSSM log-likelihood score > 0).

2.2 The AbNatiV model

2.2.1 Vector-quantized variational auto-encoder (VQ-VAE) architecture

The AbNatiV model takes aligned antibody sequences of length 149 as input, and one-hot encodes each into a tensor of dimension 149x21. Each position is represented by a vector of size 21 consisting of zeros and a one at the alphabet index of the residue under scrutiny (20 standard amino acids and a gap token).

The architecture of the models is based on a vector-quantized variational auto-encoder (VQ-VAE) framework [36], which involves a VAE with a discretisation of the dense latent space through code-vectors (**Fig. 1A**). The sequence input $x \in \{0,1\}^{149\times21}$ is first encoded into a compressed sequence representation $z_e(x) \in \mathbb{R}^{l\times d_c}$, where l represents the compressed sequence length and d_c the dimension of the code-vectors. In order to discretise $z_e(x)$ in the latent space, a learnable codebook of N code-vectors $\{e_k\}_{k=1}^{N} \subset \mathbb{R}^{d_c}$ is used. A nearest neighbour lookup is applied, so that each component $\{z_e(x)_i\}_{i=1}^{l} \subset \mathbb{R}^{d_c}$ is substituted by the closest code-vector of the codebook, resulting in the quantised embedding $z_q(x) \subset \mathbb{R}^{l\times d_c}$. Finally, $z_q(x)$ is decoded to generate the reconstructed output $\hat{x} \in \{0,1\}^{149\times21}$ having the original dimensions as the original sequence input x.

For increased codebook usage (i.e., higher perplexity), the N code-vectors are initialized with the N k-means centroids of the first training batch, and code-vectors not assigned for multiple batches are replaced by randomly sampling the current batch as detailed in Ref. [37], where a vector quantizer was applied to sound compression. In addition, the code-vectors $\{e_k\}_{k=1}^{N}$ and the encoded inputs $z_e(x)$ are l_2-normalised. The Euclidean distance of the l_2-normalised vectors is used during the nearest neighbour lookup resulting in a cosine similarity search as proposed in the image modelling model ViT-VQGAN [38]. Furthermore, the code-vectors from the codebook are updated during training by exponential moving average (EMA) with a decay of 0.9 to assure a more stable training [39].

The encoder and decoder layers are illustrated in **Figure 1B**. In the encoder, the input sequence is embedded by a patch convolutional layer [38]. A 1D-convolution layer with a kernel size K equals to its stride S embeds each of the non-overlapping patches of dimension Kx21 into a single vector of size d_{emb} (i.e., the number of channels of the 1D-convolution layer). A minimal padding was added to the

sequence input beforehand to avoid missing any sequence region. For instance, in the VHH model, with $K = S = 8$, a padding of 3 is added to compress the sequence inputs into $l = 19$ embedding vectors of size d_{emb}. Then, a sinusoidal positional encoding is added before L transformer blocks. The transformer blocks are designed as in BERT [40], with H heads in the multi-head attentions layer and a hidden dimension d_{ff} in the feed forward layer. Before quantisation, a linear layer is applied to reduce the embedding dimension d_{emb} to the size of the code-vectors d_c.

In the decoder, a linear layer is first applied to augment the dimension of the discrete embedding $z_q(x)$ to d_{emb}. Mirroring the encoder, a positional encoding is applied before L transformer blocks with the same hyper-parameters of the encoder. Ultimately, a transpose 1D-convolution layer with a softmax activation function is applied to reconstruct back the tensor into the same dimension of the original sequence inputs. All hyper-parameters were manually tuned for every model and their values are given in **Table S2**.

2.2.2 Unsupervised masked learning

The models are trained to minimize a loss function consisting of three terms:

$$loss = \| x - \hat{x} \|_2^2 + \| sg(z_e(x)) - z_q(x) \|_2^2 + \beta \| z_e(x) - sg(z_q(x)) \|_2^2$$

The first term is a reconstruction mean-squared error (MSE) between the input x and the decoder output $\hat{x}$. The second and third terms are associated with the variational autoencoder and the vector quantisation step in the latent space. Both terms are MSEs between the encoded input $z_e(x)$ and the quantised latent embedding $z_q(x)$. In particular, in the second term, stop gradient sg is applied to $z_e(x)$ to detach it from the computational graph, updating only the codebook during back propagation. In the third term, $z_q(x)$ is conversely ignored during back propagation, which drives the encoder to commit to the codebook vectors. The relative learning speed between these two terms is imposed by the scaling factor β.

The neural network is implemented using PyTorch.1.14 [41] and enhanced by the PyTorchLightning.0.7 module. The models are trained with a batch size of 128 by the Adam optimizer [42] with a learning rate of 4e-05. During training, a masking is applied to the one-hot encoded inputs. As in the training of the language transformer model BERT [40], a percentage of positions p_{mask} is selected for masking. Among these selected positions, 80% are replaced by the uniform vector of size 21 with a probability of 1/21 for each residue, which we use as a mask token. 10% are randomly replaced by another residue or gap. 10% remain unchanged so that the model does not learn to expect a fixed number of masked residues (as all sequences are aligned to 149 positions).

2.3 Antibody nativeness definition

The concept of antibody nativeness is intuitively understood as the extent to which a given sequence resembles those of native antibodies, that is of antibodies derived from the immune system under scrutiny (in this work human or camelid immune systems). Here, we provide a quantitative definition of nativeness as:

$$AbNatiV\ Nativeness = \frac{0.8 - 1}{T_R - 1} \left(\exp\left(- \frac{\sum_{i=1}^{149} \frac{1}{21} \| \hat{x}_i - x_i \|_2^2}{Sequence\ Length} \right) - 1 \right) + 1$$

where $\| \hat{x}_i - x_i \|_2^2$ is the MSE at sequence position i between the aligned input sequence x and the reconstructed output sequence $\hat{x}$ of a trained AbNatiV model. This MSE is summed over all 149 positions of the aligned sequence and normalised by the length of the input sequence (i.e., without considering the gaps opened by the alignment). As this operation gives a number X that in principle ranges in $[0, +\infty[$, where 0 would correspond to a fully native sequence that is perfectly reconstructed, we apply the function $Y = \exp(-X)$. This way, Y is now a number in $[0,1]$, where 1 means fully native, thus providing a more intuitive ranking for high and low nativeness. We wish to point out that, for typical antibody sequences from any species, the average MSE X is typically a very small number in all the models that we trained. Therefore, in this relevant range of X, $Y = \exp(-X)$ is effectively approximated by a simpler linear transformation $Y = 1 - X$ meaning that the distance between different antibody sequences is only minimally affected by the exponential transformation. Finally, the operation $(0.8 - 1) * (Y - 1)/(T_R - 1) + 1$ linearly rescales the scores so that the final nativeness score becomes a quantity directly and intuitively interpretable as an absolute value for a single sequence, and not just usable to rank different sequences (**Fig. S3**). T_R is specific to each trained model, and it denotes the optimal threshold of Y that best separates native sequences (positives in the classification) from non-native sequences (negatives in the classification). This linear transformation rescales the values of Y so that this threshold on the final nativeness score becomes 0.8 for every model. In other words, this means that a nativeness score greater than 0.8 denotes a sequence classified as native, while a score below 0.8 one classified as non-native. T_R is calculated for each trained model as follows. The precision-recall (PR) curves are generated between human sequences (Human Test & Human Diverse datasets) as positives, and non-human sequences (Mouse) as negatives for the VH, VKappa and VLambda models. Similarly, the PR curve is also calculated between VHH sequences (Camelid Test) as positives, and non-VHH sequences (Human Test and Mouse) as negatives, all computed on the $Y = \exp(-X)$ scored sequences (**Fig. S3A, S3D, S3G,** and **S3J**). For every model, the PR optimal threshold value T_R is extracted as the point closest to (1,1) (**Fig. S3B, S3E, S3H,** and **S3K**, $T_R(VH) = 0.988477$, $T_R(VKappa) = 0.992744$, $T_R(VLambda) = 0.985598$, and $T_R(VHH) = 0.990896$). The scores are thus linearly rescaled to shift T_R to 0.8 to return a final value $\in]-\infty, 1]$ for any input Fv sequence (**Fig. S3C, S3F, S3I,** and **S3L**). Not only does this rescaling make the nativeness scores from different models interpretable in the same way, but it also future proofs the definition of nativeness. The values of Tr will change if, in the future, the model is retrained on a larger or more diverse dataset, or if the architecture is further improved. However, the interpretation of the final nativeness score, which is what users will rely on, will be the same. We define AbNatiV humanness score the nativeness from AbNatiV trained on VH, VKappa, and VLambda human sequences, and AbNatiV VHH-nativeness score, that from AbNatiV trained on single-domain VHH sequences.

In addition, residue level scoring profiles are defined by applying $Y = \exp(-X)$ to the MSE reconstruction error at each position of the given sequence.

2.4 Performance metrics

All the performance metrics reported are computed by analysing 10,000 scored sequences for each database (**Table S1**). For datasets smaller than 10,000, the whole dataset is used.

2.4.1 Classification

The area under the curve (AUC) of the receiver operating characteristic (ROC) and of the precision-recall (PR) curves are computed to quantify the ability of a model to classify sequences. For ROC

curves, the AUC is equal to 1 when the classification is perfect. It is equal to 0.5 when the model performs as poorly as a classifier that is randomly sampling from a uniform distribution. For PR curves, the AUC is also equal to 1 when the classification is perfect, while it is equal to the ratio of positive entries over the total number of entries in the datasets when the classification is random.

2.4.2 The amino acid reconstruction accuracy

The amino acid reconstruction accuracy quantifies the ability of a model to reconstruct the initial unmasked input from the embedded vector of the latent space. The reconstructed outputs of the model have for each position a probability distribution over the alphabet. For each position, the most probable amino acid is selected. The amino acid reconstruction accuracy corresponds to the ratio of correctly predicted residues for every position over the length of the sequence. It is equal to 1 if all residues have been correctly reconstructed, and 0 if not even one has. It can be expressed, as follows:

$$Reconstruction\ Accuracy = \frac{\sum_{i=1}^{149} 1_{x_i=\hat{x}_i}}{149}$$

where x_i and $\hat{x}_i$ are residue at the position i of respectively the input x and the reconstructed output $\hat{x}$ of the model.

2.4.3 Benchmarking with other assessments from the literature

The GitHub implementation (https://github.com/Merck/BioPhi) of Sapiens and OASis from Biophi [43] is used to score our testing databases. The relaxed stringency level is used for the OASis assessment. The OASis score is not position discrete; hence it cannot be used for the amino acid reconstruction task.

The AbLSTM model is originally trained on its own database of human heavy chains [44]. Here, we used the pre-trained model in benchmarking, and we also retrained the AbLSTM for 10 epochs from scratch on the same single-domain, and human heavy, lambda and kappa databases used for the training of our VQ-VAE models. In the case of human VH we carried out the benchmark with both retrained AbLSTM and original pre-trained one as downloaded from https://github.com/vkola-lab/peds2019. The original hyperparameters of AbLSTM were used (embedding dimension = 64, hidden dimension= 64, batch size = 128, and learning rate = 2e-03). The loss of the AbLSTM model was employed as its humanness or VHH-nativeness scores as done in the original work [44].

2.5 Predictions on antibody therapeutics

549 antibody therapeutics from the IMGT database [45] were obtained from the BioPhi dataset [43]. This dataset includes 196 fully human therapeutic sequences and 353 therapeutics of non-human origin (mouse, chimeric, and humanised). The AUC of ROC and PR curves are computed to quantify the ability of the models to separate these two groups of sequences.

Similarly, 216 antibody therapeutics with their immunogenicity scores – expressed as the percentage of patients who developed an anti-drug-antibody (ADA) response during clinical trials – were also obtained from the BioPhi dataset [43]. These sequences were used to quantify the extent of correlation between the models nativeness scores and the observed ADA response, using the Pearson correlation coefficient and its associated p-value. For each therapeutic, the mean between the scores of VH and VL domain is used as an overall nativeness.

The humanness scores from different methods developed to humanize antibodies with which we compare our approach were obtained as computed by the authors of BioPhi and deposited in their GitHub (https://github.com/Merck/BioPhi) and in the tables of Ref. [43]. The alternative methods considered in this work are the BioPhi germline content [43], HumAb [46], IgReconstruct [47], AbLSTM [44], T20 [48], and Z-score [49] assessments. Light-chain-only antibodies (i.e., Istiratumab, Lulizumab Pegol, Placulumab and Tibulizumab) are removed from the IMGT BioPhi parsed dataset as the original pretrained AbLSTM can only scores heavy chains. Because the Fv sequence of Pexeluzimab has missing C-term residues, it is also removed from the ADA dataset and excluded from further analysis. All these sequences with their associated scores are available in **Supplementary Datasets 1** and **2**.

2.6 Grafting assessment on nanobodies

In Ref. [50] all three CDRs of 6 nanobodies were grafted onto a camelid VHH framework sequence, referred to as universal framework (UF). Binding K_D and conformational stability ΔG were experimentally measured for all six wild type (WT) nanobodies, and corresponding variants with CDRs grafted onto the universal framework. Here, we compute the nativeness scores of the 6 pairs of WT and grafted nanobodies. As the UF has intrinsically better nativeness because of its ideal framework, to understand whether our model predicts the CDRs to be in the right context or not, we compute the VHH-nativeness CDRs scores. These are defined as the sum of the MSE reconstruction scores of all residues at the CDR positions (according to the AHo numbering scheme) normalised by the length of these CDRs without gaps. $Y = \exp(-X)$ is applied to the resulting sum X to give a more interpretable number in $[0,1]$. A nativeness prediction of a CDR context is considered correct when the VHH-nativeness CDRs score of the WT nanobody is higher than that of its UF-grafted counterpart, as reflected by the experimentally measured change in binding K_D which is typically worse for the UF-grafted variant (**Fig. 4D**).

We also carried out this assessment on a much bigger scale, by computationally grafting all CDRs of 5,000 different Nbs from the Camelid Test dataset onto the UF scaffold.

2.7 Humanness assessment of nanobodies

For the analysis reported in **Figure 5**, 300 VH human sequences and 300 camelid sequences from the test datasets are scored both with the AbNatiV human heavy and camelid heavy models to provide background distributions. Then, we further scored 8 WT nanobodies from a SARS-CoV-2 study [51] and their humanised counterpart as reported in Ref. [52], and 3 therapeutic nanobody sequences (Envafolimab, Caplacizumab, and Rimteravimab) available from the therapeutic database Thera-SAbDab [53].

3 Results

3.1 The AbNatiV model

AbNatiV is a deep learning model trained on immune-system-derived antibody sequences. It employs an architecture inspired by that of the vector-quantized variational auto-encoder (VQ-VAE), originally proposed for image processing (i.e., for tensors of rank 3) [36]. The AbNatiV architecture compresses amino-acid sequences (encoded as tensors of rank 2) into a bottleneck layer, also called embedding, where each latent variable is mapped to the closest code-vector from a learnable codebook, prior to reconstruction with a decoder (**Fig. 1A**). This vector quantisation from the codebook leads to a discrete latent representation rather than a continuous one as in standard VAE. This VQ architecture was chosen because protein sequences are discrete objects and thus may favour a discrete representation, and

because it was shown to circumvent issues of posterior collapse that sometimes affect standard VAEs [36]. Our model contains in the encoder and decoder both patch convolutional layers and transformers (**Fig. 1B**). These are respectively more suitable to capture local interactions along the sequence (i.e., local motifs), and long-range interactions between such local motifs or individual residues, which may be mediated by tertiary contacts. High codebook usage (i.e., high perplexity) is ensured in the bottleneck by a k-means initialisation of the codebook and a cosine similarity search during the nearest neighbour lookup quantisation, to prevent poor data representation and maintain a robust training [54] (see Methods and **Fig. S4**).

The model is trained with masked unsupervised learning. Unsupervised learning works on the assumption that every sequenced antibody follows some set of biophysical and evolutionary rules that allow it to be produced by organisms and to carry out its biological function without causing toxicity. AbNatiV is built to impose a bottleneck in the network that forces a compressed representation of the input sequence, which is then reconstructed by the decoder. If the amino acids within the input sequences were fully independent from each other, this compression and subsequent reconstruction would be impossible. However, if some structure exists in the input, as it is the case for antibody sequences, this structure should be learnt and consequently leveraged when forcing the input through the network bottleneck. Therefore, the AbNatiV architecture is in principle capable of learning a representation of natural antibodies that captures high-order relationships between residue positions to provide a highly sensitive measure of antibody nativeness. To ensure that the model learns meaningful high-order relationships, we also employed masked learning. During training the input sequence is masked by removing information on the identity of a random subset of residues, and the training task is to reconstruct the full sequence, including correctly predicting the identity of the masked residues (see Methods). Taken together, AbNatiV architecture and masked unsupervised learning strategy drive the model to capture the essential features that are common across a database of native antibody sequences.

AbNatiV is trained on aligned sequences of native antibody from curated immune repertoires from the OAS database [33] and other sources (see Methods). The model is trained separately on human VH, VKappa, Vlambda, and camelid VHH sequences (~2 million unique sequences in each training set). The Kappa and Lambda light chains are treated separately due to their significant differences. AbNatiV takes around 1 hour per epoch to train (~17 million parameters) on a single GPU (NVIDIA RTX 8000). For each model, a validation dataset of 50,000 unique sequences different from those in the training set monitors the absence of overfitting (**Fig. S5**) and is used for hyperparameter optimisation. 10,000 further unique sequences, distinct from those in training and validation sets, are kept aside for testing.

For each input Fv sequence, the trained AbNatiV models return an antibody nativeness score and a sequence profile. The nativeness score quantifies how close the input sequence is to the learned distribution, that is to a native antibody sequence derived from the immune system the model was trained on (human or Camelid in this work). To facilitate the interpretation of this score and the comparison of scores from the different trained models, the AbNatiV score is defined in such a way that it approaches 1 for highly native sequences, and that 0.8 represents the threshold that best separates native and non-native sequences (see Methods). In the case of AbNatiV trained on VH, VLambda and VKappa human chains, this score is referred as to the AbNatiV humanness score (**Fig. 1C**). Similarly, for AbNatiV trained on VHH camelid sequences, the score is referred to as the AbNatiV VHH-nativeness score.

The sequence profile consists of one number per position in the aligned input sequence, so it contains a total of 149 entries including gaps. Here too, entries approaching 1 denote high nativeness, and smaller than 1 increasingly lower nativeness. This profile is useful to understand which sequence regions or residues contribute most to the overall nativeness of the sequence, and which may be liabilities. As an example, **Fig. 1D** shows the humanness profile of the VH sequence of a mouse antibody (WT precursor) that contains many low-scoring regions that could be immunogenic in humans, compared to that of its humanised counterpart: the therapeutic antibody Refanezumab. The profile of Refanezumab contains far fewer low-scoring regions, and these are mostly found in the CDR loops, which are of mouse origin and were grafted into a human Fv framework during humanisation (**Fig. 1D** and **S2**). This example shows that sequence profiles can be powerful tools to guide antibody engineering, by facilitating the design of mutations to improve antibody nativeness.

3.1.1 Classification of human antibodies

To quantify the performance of AbNatiV, we first assessed its ability to discriminate between human antibody Fv sequences and antibody Fv sequences from other species. The area under the receiver operating characteristic curve (ROC-AUC) and that under the precision-recall curve (PR-AUC) are used to quantify the ability of the models to correctly classify sequences (**Fig. 2** and **S6**). For example, AbNatiV can accurately distinguish the VH human sequences of its Test set from VH mouse sequences based on their humanness score distribution with a ROC-AUC of 0.996 (**Fig. 2B**) and PR-AUC of 0.997 (**Fig. 2C**). Similarly, AbNatiV can successfully discriminate between human and rhesus (monkey, *Macaca mulatta*) sequences. Despite the high genetic similarity between these two organisms, the model can separate VH sequences very well, with ROC-AUC of 0.962 and PR-AUC of 0.968 (**Fig. 2B** and **2C**).

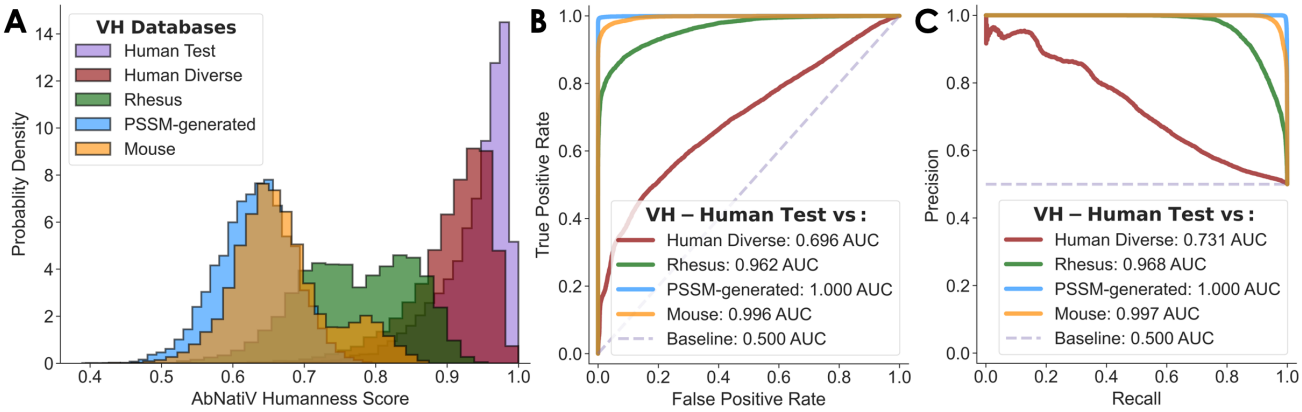

Figure 2. Performance on VH sequence classification. (**A**) The AbNatiV humanness score distributions of the Human Test (purple), Human Diverse (red), Rhesus (green), PSSM-generated (blue), and Mouse (orange) VH antibody datasets. The PSSM-generated database is made of artificial sequences randomly generated using residue positional frequencies from the PSSM of the Human Test dataset. Each dataset contains 10,000 sequences. (**B, C**) Plots of the ROC curves (**B**) and PR curves (**C**) computed to represent the ability of AbNaTiV to distinguish the VH Human Test set from the other datasets (see legend, which also reports the area under the curve). The baseline (dashed line) corresponds to the performance that a random classifier would have.

We further employed two control datasets in our benchmark: one for the learning of high-order relationships, and one to confirm the lack of overfitting and the ability of the model to generalise to unseen sequence space. For the latter, we employed a dataset of highly diverse human Fv sequences

assembled by the authors of Ref. [43], which we first ensured were not present in our Training, Validation or Test sets. As expected, we find a good overlap between the distribution of humanness scores of our Test set and that of this control dataset (**Fig. 2** and **S6**, ROC-AUC of 0.696, 0.474, 0.418 respectively for VH, VLambda, and Vkappa). This performance on the control dataset is in line with our assessment of lack of overfitting (**Fig. S7**), and it makes us confident in the ability of the model to generalise to sequences distant from those in its Training set.

As a control for the learning of high-order relationships, we generated datasets of artificial Fv sequences constructed by picking residues at random following the positional residue frequencies observed in human Fv sequences (see Methods and **Fig. S1**). We call these datasets PSSM-generated sets. If one looks at each residue position individually, these artificial sequences are indistinguishable from real human sequences, as they are constructed only using residues observed in human sequences at each position (with log-likelihood>0 and following the observed residue frequency distribution, see Methods). However, as residues at each position of the artificial sequences have been chosen independently of residues at other positions, any high-order relationship observed in these sequences should be compatible with random expectation. Remarkably, we find that AbNatiV can perfectly separate real VH Human sequences from PSSM-generated ones (AUC=1; **Fig. 2**), and that the separation is also excellent for VLambda and VKappa (PR-AUC respectively of 0.990 and 0.993; **Fig. S6**). This performance attests the ability of AbNatiV to learn complex high-order relationships observed within native human Fv sequences beyond their simple amino acid composition.

VH	**Classification (AUC) Human Test vs :**								**Reconstruction accuracy**	
	Mouse		**Rhesus**		**Human Diverse**		**PSSM-generated**		**Human Test**	**Human Diverse**
	ROC	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**		
AbNatiV	**0.996**	**0.997**	**0.962**	**0.968**	0.696	0.731	**1.000**	**1.000**	**0.962**	0.937
OASis (relaxed)	0.918	0.897	0.526	0.570	0.205	0.346	0.976	0.982	N/A	N/A
Sapiens	0.980	0.982	0.581	0.626	0.172	0.335	0.991	0.993	0.918	**0.956**
AbLSTM	0.961	0.963	0.742	0.721	0.220	0.348	0.997	0.998	0.807	0.868
AbLSTM Retrained	0.965	0.967	0.791	0.777	0.325	0.385	0.996	0.997	0.822	0.859

Table 1. Evaluation of the classification and reconstruction tasks for human VH sequences. The assessment is carried out for AbNatiV trained on human VH sequences (first row) and other computational approaches that can assess humanness (other rows). AbLSTM retrained corresponds to the AbLSTM model retrained on the same training set of AbNatiV (see Methods). The first eight columns report the area under the curve for both ROC and PR curves (shown in **Fig. S7**), assessing the ability of the models to separate sequences in the human Test set from those from mouse, rhesus, human-diverse and PSSM-generated (see column headers). The human-diverse dataset is used here as a control, as the task is to separate human sequences from other highly diverse human sequences. The last two columns quantify the ability of each model to reconstruct human sequences in each dataset (column header). The OASis method does not carry out reconstruction. Entries in bold correspond to the best performance in that column.

We then compared the performances of AbNatiV with that of other computational methods developed to assess the humanness of antibody sequences in the context of antibody humanisation (**Tables 1**, **S3**, **S4** and **Fig. S7**, **S8**, **S9**). More specifically, we focus on the recently introduced OASis 9-mer peptide similarity score [43], the Sapiens transformer model [43], and the AbLSTM model [44], as these

approaches were shown to outperform older methods. Our results show that AbNatiV outperforms all alternative approaches on all classification tasks (**Tables 1**, **S3**, **S4**). The biggest difference is observed in the Human vs. Rhesus classification, where for VH sequences the AbNatiV ROC-AUC is 0.962 while that of the best alternative method, AbLSTM, is 0.742, which increases to 0.791 once the AbLSTM architecture is re-trained on our training set (**Table 1**). Lower performances of the alternative models are also shown for the human vs. mouse and human vs. PSSM-generated classification tasks. We have not included in this benchmark the recently introduced Hu-mAb method [46], because we could only access it as a webserver that processes a single sequence per run. However, as Hu-mAb is trained with supervised learning for the specific task to distinguish between human and mouse sequences, we would expect it to do extremely well at the mouse vs. human classification task, and perhaps not as well on other tasks. We note that the Human Test set vs. Human Diverse set classification task yields ROC-AUC scores smaller than the 0.5 baseline AUC for all tested alternative methods. This result may stem from the fact that many of the sequences in the Human Diverse databases are contained in the training set of the other methods, which is certainly the case for the Sapiens transformer [43], and hence are overclassified with respect to our Human Test set. In addition, amino acid reconstruction accuracies were computed for all methods (except OASis as the method is not reconstruction-based). The reconstruction accuracy quantifies the ability of a model to reconstruct the initial input from the embedding in the latent space. Both AbNatiV and Sapiens rely on masked learning, while AbLSTM relies on standard unsupervised learning. We find that the former models have higher reconstruction accuracies than the AbLSTM model (96%, 92% and 81% on the Human Test set respectively for AbNatiV, Sapiens and AbLSTM). Sapiens reconstructs slightly better than AbNatiV the VH sequences in the Human Diverse dataset (respectively 94% and 96%). However, it should be noted that the Human Diverse dataset is contained in the training set of Sapiens [43]. Similar results are found for VKappa and VLambda lights chains, when comparing AbNatiV with the OASis and Sapiens methods (**Table S3**, **S4** and **Fig. S8**, **S9**), while the AbLSTM humanness score is not defined for light chains [44]. Curiously, in contrast with VH sequences, AbNatiV exhibits higher reconstruction accuracy than Sapiens also on the VL sequences in the Human Diverse dataset (respectively 0.984 vs 0.954 for VLambda, and 0.985 vs 0.963 for VKappa). Taken together, these results demonstrate that AbNatiV is a precise humanness assessment method that has learned high-order relationship between residues to identify antibody sequences derived from human immune systems.

3.1.2 Application to antibody therapeutics

The assessment of humanness is a critical step of antibody drug development, with the goal of ensuring that drug candidates have minimal risk for administration to patients. Therefore, we ran AbNatiV on therapeutic antibody sequences and averaged the humanness score of the heavy and light chains from the relevant AbNatiV model (i.e., trained either on VH, VKappa, or VLambda, see Methods). More specifically, we evaluated the performance of the method on distinguishing 196 human therapeutics from 353 antibodies therapeutics of non-human origin (mouse, chimeric, and humanised). The precision-recall (PR) curve (**Fig. 3A**) and ROC curve (**Fig. S10**) are computed for AbNatiV and 7 other computational approaches (see Methods and **Table 2**). AbNatiV outperforms all other methods with a PR-AUC of 0.975 and a ROC-AUC of 0.981. The second-best methods after AbNativ are OASis with a PR-AUC of 0.963 and Hu-mAb with a ROC-AUC of 0.979.

A central interest in humanisation of antibodies is to reduce their immunogenicity in human immune systems. One way to assess immunogenicity in early-stage clinical trials is to assess the number of patients that develop anti-drug antibodies (ADAs) in response to the administration of therapeutic antibodies [55]. We find that the AbNatiV humanness score (i.e., the average of the AbNatiV humanness scores of the VH and VL, see Methods) shows a Pearson correlation coefficient (R) of -

0.50 (p-value ~ $9x10^{-15}$) with the percent of patients that developed ADAs upon treatment, which is available for 216 different therapeutic antibodies (**Fig. 3B**). We note that these ADA data are highly heterogeneous and therefore there is no reason to expect much stronger correlations. The percent of patients who developed an ADA response is determined in different studies carried out in very different ways. In particular, the dosage of the therapeutic antibody candidate and the length of the study (i.e., the number of doses administered and the total study time) can vary widely among different therapeutic candidates. It is therefore foreseeable that a highly immunogenic antibody which is administered only once and at a relatively low dose would elicit a weaker ADA response than a less immunogenic antibody that is administered at high dose for an extended period. The reason for these discrepancies is that these clinical studies are designed around the specific requirements of the drug candidate under scrutiny, rather than to quantitatively compare the immunogenicity of different drug candidates.

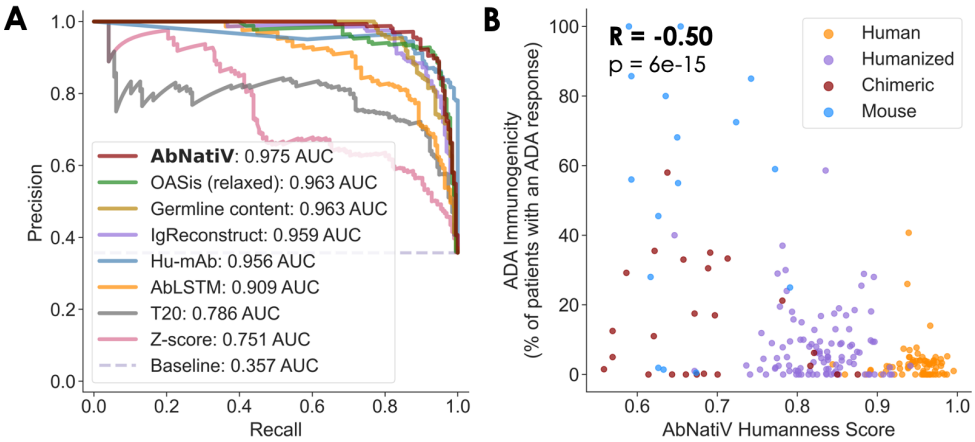

Figure 3. Performance on antibody therapeutics. (**A**) Plot of the PR curves of the classification of 196 human-derived therapeutics from 353 therapeutics of non-human origin (mouse, chimeric, and humanised) carried out with AbNatiV (in red) and seven other computational methods (see legend, which also reports the AUC values). The baseline (dashed line) corresponds to the performance expected from a random classifier. Corresponding ROC curves can be found in **Fig. S10**. (**B**) Scatter plot of the AbNatiV humanness score of 126 antibody therapeutics and their ADA immunogenicity score, expressed as the percentage of patients developing an ADA response in each study. The Pearson correlation (R) and p-value are reported on top left corner. Sequences are coloured based on their origin (i.e., human in orange, humanised in purple, chimeric in red, and mouse in blue).

Method	Human vs Non-Human	
	ROC AUC	PR AUC
AbNatiV	**0.981**	**0.975**
OASis (relaxed)	0.975	0.963
Germline content	0.971	0.963
IgReconstruct	0.971	0.959
Hu-mAb	0.979	0.956
AbLSTM	0.937	0.909
T20	0.898	0.786
Z-score	0.837	0.751

Table 2. Performance on the classification of antibody therapeutics. The assessment is carried out for AbNatiV (first row) by averaging the AbNatiV humanness scores of the heavy and light chains

from the relevant AbNatiV model (i.e., trained either on VH, VKappa, or VLambda, see Methods), and for other computational methods (table rows). The classification task consists in distinguishing 196 human-derived therapeutic antibodies from 353 therapeutic antibodies from a different origin (mouse, chimeric, and humanised). The area under the curve for both ROC and PR curves are reported in the first two columns. Entries in bold correspond to the best performance in that column.

3.1.3 Classification of native camelid nanobodies

The development of single-domain antibodies has been gathering even more momentum since the approval of Caplacizumab in 2019, the first nanobody-based therapeutic [6]. Nanobodies (VHHs) are naturally expressed in camelids and can exhibit advantageous stability and solubility properties combined with a small size that allows for better tissue penetration, while retaining the affinity and specificity of full-length antibodies [5]. When trained on VHH sequences, AbNatiV returns a VHH-nativeness score that quantifies the resemblance of antibody sequences to native camelid nanobodies, and hence the ability of a VH sequence to fold independently of a VL counterpart.

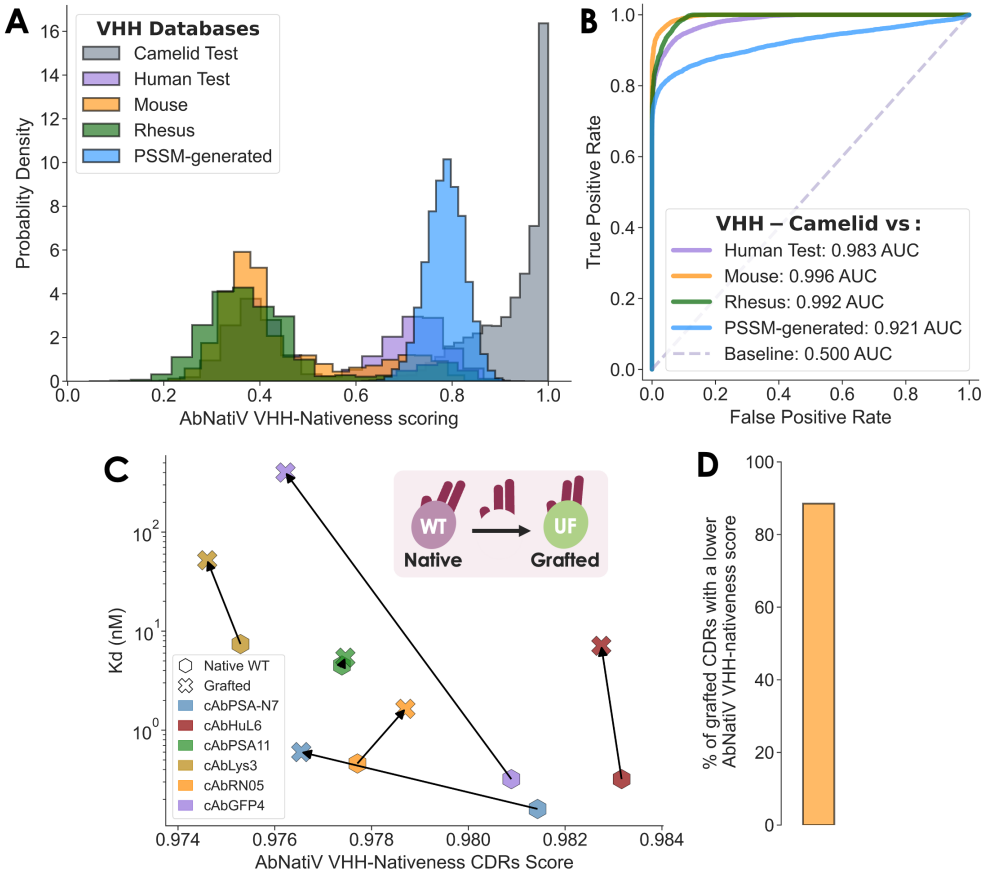

Figure 4. Performance on VHH sequences derived from camelids. (**A**) The AbNatiV VHH-nativeness score distributions of the VHH Camelid Test (in pink), VH Human Test (in purple), VH Mouse (in orange), VH Rhesus (in green), and VHH PSSM-generated (in blue) antibody datasets. The VHH PSSM-generated database is made of artificial sequences randomly generated using residue positional frequencies from the PSSM of the Camelid Test dataset. Each dataset contains 10,000 sequences. (**B**) Plot of the ROC curves used to quantify the ability of AbNaTiV to distinguish the VHH Camelid Test set from the other datasets (see legend, which also reports the AUC values). The baseline (dashed line) corresponds to the performance expected from a random classifier. The corresponding

PR curves are given in **Fig. S12**. (**C**) Plot of the binding K_D, reported in Ref. [50], as a function of the AbNatiV VHH-nativeness score computed across all CDR positions of 6 nanobodies (see legend) before and after grafting of all three CDRs onto a camelid universal framework (UF). An arrow is directed from the native sequence in the WT framework to the grafted one. (**D**) 89% of all three CDRs among a test set of 5,000 nanobodies has a lower AbNatiV VHH-nativeness score when they are computationally grafted onto the camelid UF (see Methods) than when they are in their native scaffold.

We find that AbNatiV accurately discriminates VHH sequences from the VH sequences of human (0.983 ROC-AUC), mouse (0.996), and rhesus (0.992) (**Fig. 4A, 4B** and **S12**). The ROC-AUC between PSSM-generated artificial VHH sequences and real camelid VHH sequences is 0.921. To the best of our knowledge, AbNatiV is the first approach to quantify the nativeness of nanobodies. Therefore, to compare with a different model, we retrained the AbLSTM architecture, originally developed for human VH sequences, on our nanobody training set (see Methods). We find that the AbNatiV model shows higher classification performance than the AbLSTM retrained model on all tasks except for the PSSM-generated classification (**Table 3** and **Figure S13**).

VHH	Classification (AUC) Camelid Test vs :								Reconstruction accuracy
	Human Test		Mouse		Rhesus		PSSM-generated		Camelid Test
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	
AbNatiV	**0.983**	**0.985**	**0.996**	**0.996**	**0.992**	**0.992**	0.921	0.947	**0.959**
AbLSTM retrained	0.966	0.968	0.992	0.992	0.988	0.988	**0.952**	**0.960**	0.862

Table 3. Evaluation of the classification and reconstruction tasks for camelid VHH sequences. The assessment is carried out for AbNatiV trained on camelid VHH sequences (first row) and the AbLSTM model retrained on the same training set of AbNatiV (see Methods and second row). The first eight columns report the area under the curve for both ROC and PR curves, assessing the ability of the models to separate sequences in the camelid Test set from those from human, mouse, rhesus, and PSSM-generated (see column headers). The last column quantified the ability of each model to reconstruct camelid sequences from the camelid Test set. Entries in bold correspond to the best performance in that column.

3.1.4 CDR nativeness for grafting experiment

The grafting of target specific CDRs onto a different framework scaffold is a common technique to design antibody with enhanced properties (e.g., low immunogenicity, higher stability or expressibility, etc.) [50,56,57]. In the case of nanobodies, a specific camelid framework, referred to as universal framework (UF), was shown to retain high conformational stability and prokaryotic expressibility almost independently of its CDR loops [50]. In that study, all three CDRs of 6 unrelated nanobodies targeting different antigens were grafted onto the UF. Binding affinity (K_D) and conformational stability (ΔG) were experimentally measured for all six wild type (WT) nanobodies, and corresponding UF variants with the grafted CDRs. Upon grafting, the binding K_D worsened for most variants, probably because the CDRs now make some non-native interactions with the UF sequence, which affects their conformation and consequently antigen binding, even if the conformational stability improved upon grafting because of the superior stability of the UF [50]. AbNatiV provides a direct sequence-based approach to assess the nativeness of these CDRs within the VHH UF and their WT framework, by computing the VHH-nativeness score across all CDR positions (see Methods). We find

that for 4 of these 6 grafting examples, AbNatiV scoring anticorrelates with the experimentally measured change in binding K_D (**Fig. 4C**). Specifically, AbNatiV attributes a worse (lower) VHH-nativeness score to these four sets of CDRs when they are grafted onto the UF than when they are found in their WT framework, in agreement with the experimental measurement of a worse (higher) binding K_D. An example of the nativeness profile before and after grafting is provided in **Figure S14**. The two cases in which the VHH-nativeness score is essentially identical or slightly worse in the WT framework than it is in the UF correspond to those two cases where the experimental change in binding K_D upon grafting was the smallest (cAbRN05 and cAbPSA11 in **Fig. 4C**).

Encouraged by these findings on six experimentally characterised grafting examples, we sought to obtain more robust statistics by computationally grafting all three CDRs of 5,000 different nanobodies from the VHH Test set onto the UF scaffold. We find that in 89% of cases AbNatiV computes a lower VHH-nativeness score for the CDRs grafted in the UF than for the CDRs in their native WT framework (**Fig. 4D**). Taken together, the results of these analyses suggest that AbNatiV can accurately determine whether CDR loops are in the right context.

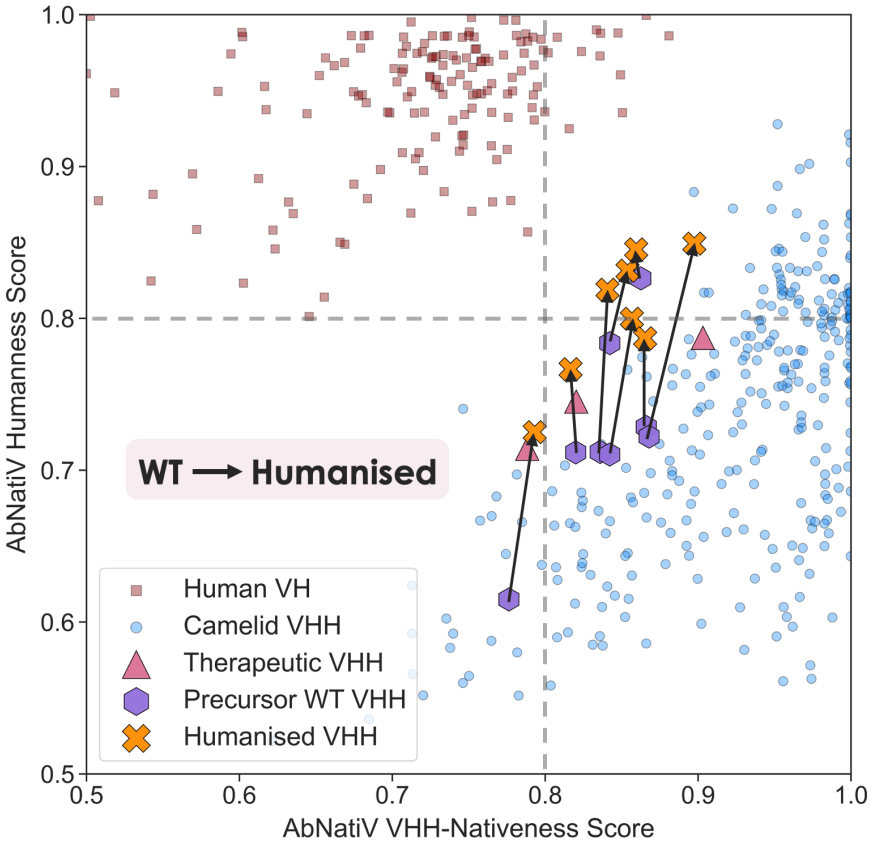

Figure 5. Combining AbNatiV humanness and VHH-nativeness. Plot of the AbNatiV humanness and VHH-nativeness scores of 300 sequences from the VH Human Test (in red), and VHH Camelid Test (in blue) datasets, along with 3 nanobody therapeutics (in pink), and 8 WT nanobodies (in purple) with their humanised counterpart (in orange) [52]. An arrow is directed from the WT sequence to the humanised one. Two dashed lines at 0.8 represent the threshold that best separates native from non-native sequences as defined in Methods. Only sequences with a score in [0.5,1] are represented for better readability.

3.2 Humanisation of nanobodies

With the recent surge of interest in the use of nanobodies as therapeutics, the humanisation of nanobodies has emerged as a crucial requirement to improve their therapeutic index and reduce immunogenicity risks for clinical applications [52,56,58]. **Figure 5** depicts the AbNatiV evaluation of the humanness and VHH-nativeness of 3 nanobody therapeutics, and of 8 WT nanobodies from a SARS-CoV-2 study [51] and their humanised counterpart [52]. To provide a reference background distribution, 300 randomly selected human VH sequences and 300 camelid VHH sequences are also plotted. The cluster of human sequences that score relatively well in VHH-nativeness derive from the IGHV-3 germline gene (**Fig. S15**), consistent with the genetic origin of natural camelid nanobodies [59].

We find that all humanised nanobody sequences are assigned an AbNatiV humanness score higher than their WT counterpart. Importantly, this improvement of humanness impacts their VHH-nativeness only weakly or not at all (**Fig. 5**). This finding is in line with the non-significant or very small change observed experimentally in the binding K_D of these nanobodies upon humanisation [52]. These small changes of binding suggest that the humanised nanobodies retained their physical and structural properties avoiding a deterioration of their nativeness in line with the AbNatiV prediction.

We note that the results in **Figure 5** provide a framework to exploit AbNatiV for the rational humanisation of nanobody sequences. For example, in a first step, liabilities in the nanobody humanness profile can be identified as in **Fig. 1D**. Then, in a second step, the human VH PSSM (**Fig. S1A**) can be used to suggest potentially humanising mutations at each of those sites. Finally, these mutations are accepted only if they improve the AbNatiV humanness score while leaving unchanged or further improving the AbNatiV VHH-nativeness score. To minimise the chances of affecting antigen binding, this procedure can be limited to the framework regions, and perhaps only to solvent-accessible framework sites, which can be determined from a structure or structural model [60,61], as commonly done in humanisation strategies based on framework resurfacing.

4 Discussion

In this work we have introduced AbNatiV, a VQ-VAE-based antibody nativeness assessment method that can evaluate the likelihood of input sequences to belong to the distribution of immune-system derived antibodies (human VH and VL domains and camelid VHHs). AbNatiV provides both an interpretable overall score for the full sequence, and a nativeness profile at the residue level, which can be exploited to guide antibody engineering and humanisation. The integration of masked and unsupervised learning with the deep VQ-VAE architecture allows AbNatiV to capture complex high-order interactions. AbNatiV successfully discriminates natural sequences from artificial sequences generated following the natural positional residue frequency, and it can distinguish human antibodies or camelid nanobodies from antibodies from other species. Compared to alternative methods developed for antibody humanisation, AbNatiV exhibits higher classification performances, while often being trained on a smaller number of sequences (~2 million) for fewer epochs (<15). To put these numbers in context, the deep VH transformer model Sapiens was trained on 20 million sequences for 700 epochs [43].

Furthermore, AbNatiV exhibits increased performance over alternative methods when classifying human-derived antibody therapeutics from therapeutic antibodies of non-human origin, which also reflects the robustness of the AbNatiV assessment beyond the span of its training and test sets. We have further shown that AbNatiV humanness score have a statistically significant correlation (R= - 0.5) with the percent of patients that developed ADA in clinical studies. This evaluation of

immunogenicity with the ADA database is commonly employed to benchmark immunogenicity assessments methods [43,46,62], and therefore we also performed it in our work. However, these ADA data exhibit a substantial level of heterogeneity, as the database was assembled using immunogenicity data from different clinical studies reported in the literature, with experimental conditions (e.g., number of patients, dosage, study length) varying substantially among studies. As an example, Basiliximab was tested on 339 patients (https://www.ema.europa.eu/), while Disitamab only on 58 [63]. In the study considered in the ADA dataset that we used, Disitamab is reported to elicit an ADA response in 58.6% of the patients [63]. However, in a more recent publication on a larger study with a more uniform design (80 patients with the same dosage instead of 58 patients with 4 different dosages), Disitamab was shown to elicit ADA response in 23.8% of the participants [64], which is less than half of the number previously estimated. This example shows that the degree of heterogeneity of this ADA database should be considered when expecting quantitative correlations with immunogenicity predictions. Nevertheless, a recently introduce method, called Hu-mAb, showed a slightly better correlation with these ADA data (R= - 0.58) [46]. Hu-mAb is a random-forest classifier trained in a supervised way to differentiate human from mouse sequences. As supervised learning is well known to typically outperform unsupervised learning, and as the ADA dataset contains only human, mouse, chimeric, or humanised antibodies from mouse precursors, it is perhaps not surprising that a supervised learning approach specifically trained to separate mouse from human antibodies shows a slightly stronger correlation with these data. In this work, we chose to develop a model trained with unsupervised learning because we want it to be applicable to any input Fv sequence, as opposed to just mouse and human sequences. One of the main reasons we developed AbNatiV is to use it in synergy with emerging approaches of de novo antibody design, which typically yield artificial sequences whose latent distribution may be specific to the design method employed.

We also trained AbNatiV to quantify the nativeness of nanobodies. The resulting model exhibits very high classification performance in distinguishing VHH sequences derived from camelids from VH sequences from other species and from PSSM-generated artificial VHH sequences. The ability to discriminate artificial sequences confirms that the correct classification of VHHs does not solely rely on the presence of nanobody hallmark residues [56], as these are also present in the artificial PSSM-generated VHH sequences. However, while the discrimination performance of native nanobody sequences from artificial ones is excellent, it is not as good as that of AbNatiV trained on human sequences (ROC-AUC of VHH: 0.921, VH: 1.000, VKappa: 0.991, and VLambda: 0.987). This observation may suggest that a bigger, and especially more diverse, VHH training dataset could be beneficial. While AbNatiV-VHH is trained on the same number of sequences of AbNatiV-Humanness, these come from a much more restricted number of studies. Therefore, our VHH dataset has more limited diversity that the human one and it also comprises nanobodies from different camelid species (llamas, dromedaries, vicugna, etc.; **Table S1**), which may slightly confuse the model and demand for a larger training dataset. Quite generally, we expect that the publication of additional camelid immune repertoires will be beneficial for data-driven approaches like AbNatiV, which have the potential to facilitate and accelerate nanobody development and humanisation.

We have also shown that AbNatiV can be used to assess whether CDR loops are in the right context or not (**Fig. 4C,D)**. This observation demonstrates the ability of the model to capture long-range interactions between CDRs and framework regions and shows that AbNatiV can assist CDR grafting. For example, the CDR nativeness loss calculated by AbNatiV is consistent with the experimentally observed loss of binding affinity upon CDR grafting in a different framework (**Fig. 4C**). Yet, a quantitative correlation with the magnitude of the change in K_D is not observed, most likely because only a subset of non-ideal CDR-framework contacts resulting from grafting actually translates to an affinity loss, in a way that is highly specific to the nanobody-antigen binding pose. We envisage

that these applications of AbNatiV may increase the effectiveness and success of *de novo* antibody design methods based on the grafting of designed CDR loops [19,20,65]. We have focussed our analysis on VHH sequences. However, the exact same approach can be carried out with AbNatiV-humanness to select human scaffold sequences that serve as better receptors for CDR grafting from non-human sources, such as murine CDRs, designed CDRs, or CDRs from a synthetic library as we showcased in **Fig. 1D**.

Nanobodies exhibit significant structural differences from human VH domains that enable them to fold independently of a VL counterpart. For instance, the CDR3 of nanobodies is often longer and sometimes folds back to interact with the framework [5,52]. During the process of humanisation for therapeutic purposes, it is crucial to improve humanness while preserving these traits, as they translate into high stability and binding affinity. As shown in **Figure 5**, AbNatiV provides an effective sequence-based way to monitor concurrently the humanness and the VHH-nativeness of a nanobody throughout the humanisation process. As users can also look at the residue profiles, the assessment method proposed by AbNatiV is granular and interpretable, allowing for an in-depth analysis of the expected impact of humanisation on the VHH-nativeness of nanobodies, which if decreased can lead to binding affinity losses. This comprehensive approach will facilitate informed decisions in therapeutic nanobody development.

Finally, we note that the trained AbNatiV models may facilitate applications of semi-supervised learning, even if we have not explored this avenue in this work. Semi-supervised learning, also known as low-N learning, combines a small amount of labelled data with a large amount of unlabelled data during training [66–68]. The embedding of the VQ-VAE, and possibly also the last hidden layer of the decoder, can be seen as an effective way to distil the fundamental features of antibody variable domains into a representation that is semantically rich and structurally, evolutionarily, and biophysically grounded [69]. The compactness of this representation, and the fact that it was built by learning from a large number of functional sequences, means it can be used as input to train a supervised model (top-model) with few free parameters, which therefore may be expected to generalise with relatively few labelled training data [67]. Approaches of semi-supervised learning with protein directed-evolution data have recently been successfully deployed and were shown to be able to generalize to unseen regions of sequence space [66,68,70].

In summary, we expect that AbNatiV will facilitate antibody and nanobody development, as it provides a sequence-based, highly accurate, and interpretable way to quantify humanness and VHH-nativeness from the knowledge of the sequence alone. Looking into the future, it is reasonable to expect that computational approaches of de novo antibody design will be increasingly adopted to generate novel antibodies. In this context, AbNatiV provides a holistic way to select the best designed antibodies or nanobodies to target epitopes of interest, for instance by ensuring high humanness or by facilitating the selection of a framework highly compatible with designed CDR loops. Antibodies designed in this way will have high nativeness, and therefore can be expected to share similar specificities and in vivo properties as immune-system-derived antibodies. Besides low immunogenicity, these properties include favourable half-life and low self-antigen cross-reactivity, which are essential for successful clinical development. Overall, we believe that approaches like AbNatiV will constitute a step-change in our ability to design de novo antibodies with in vivo properties highly competitive with those of antibodies isolated from immune systems.

5 Conflict of Interest

The authors declare no conflict of interest.

6 Author Contributions

PS conceived and supervised the project. AR developed the deep learning architecture with the guidance of MG and PS. AR parsed and collected the training data with the help of AS and KD. AR and PS critically analysed the results and wrote the first version of the paper. All authors analysed data and edited the paper.

7 Funding

P.S. is a Royal Society University Research Fellow (URF\R1\201461). We acknowledge funding from UKRI EPSRC (EP/X024733/1) and from an Isaac Newton Trust/Wellcome Trust ISSF/University of Cambridge Joint Research Grant (MBAG/624 RG89305).

8 Data Availability Statement

All data needed to evaluate the conclusions in this article, or that are necessary to interpret, verify and extend the research in the article are present in the paper and/or the Supplementary Materials. Additional details are available from the corresponding author on request.

9 Code Availability Statement

The AbNatiV code repository including the trained models is available at https://gitlab.developers.cam.ac.uk/ch/sormanni/abnativ. The method is also made available as a webserver at www-cohsoftware.ch.cam.ac.uk/index.php/abnativ. To access the server, users need to register a free account and log in.

10 References

1. Goldman, R. Antibodies: indispensable tools for biomedical research. *Trends Biochem. Sci.* **25**, 593–595 (2000).
2. Trier, N. H. & Houen, G. Antibodies as Diagnostic Targets and as Reagents for Diagnostics. *Antibodies* **9**, 15 (2020).
3. Kaplon, H., Crescioli, S., Chenoweth, A., Visweswaraiah, J. & Reichert, J. M. Antibodies to watch in 2023. *mAbs* **15**, 2153410 (2023).
4. Hamers-Casterman, C. *et al.* Naturally occurring antibodies devoid of light chains. *Nature* **363**, 446–448 (1993).
5. Muyldermans, S. Nanobodies: Natural Single-Domain Antibodies. *Annu. Rev. Biochem.* **82**, 775–797 (2013).
6. Scully, M. *et al.* Caplacizumab Treatment for Acquired Thrombotic Thrombocytopenic Purpura. *N. Engl. J. Med.* **380**, 335–346 (2019).
7. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495–497 (1975).
8. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554 (1990).
9. Sormanni, P., Aprile, F. A. & Vendruscolo, M. Third generation antibody discovery methods: in silico rational design. *Chem. Soc. Rev.* **47**, 9137–9157 (2018).
10. Sellés Vidal, L., Isalan, M., Heap, J. T. & Ledesma-Amaro, R. A primer to directed evolution: current methodologies and future directions. *RSC Chem. Biol.* 10.1039.D2CB00231K (2023) doi:10.1039/D2CB00231K.

11. Winter, G., Griffiths, A. D., Hawkins, R. E. & Hoogenboom, H. R. Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12**, 433–455 (1994).
12. Murphy, A. J. *et al.* Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. *Proc. Natl. Acad. Sci.* **111**, 5153–5158 (2014).
13. Lee, E.-C. *et al.* Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nat. Biotechnol.* **32**, 356–363 (2014).
14. Traggiai, E. *et al.* An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. *Nat. Med.* **10**, 871–875 (2004).
15. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453**, 667–671 (2008).
16. Lonberg, N. Fully human antibodies from transgenic mouse and phage display platforms. *Curr. Opin. Immunol.* **20**, 450–459 (2008).
17. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.* **114**, 944–949 (2017).
18. Lu, R.-M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* **27**, 1 (2020).
19. Sormanni, P., Aprile, F. A. & Vendruscolo, M. Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. *Proc. Natl. Acad. Sci.* **112**, 9902–9907 (2015).
20. Aguilar Rangel, M. *et al.* Fragment-based computational design of antibodies targeting structured epitopes. *Sci. Adv.* **8**, eabp9540 (2022).
21. Aprile, F. A. *et al.* Selective targeting of primary and secondary nucleation pathways in Aβ42 aggregation using a rational antibody scanning method. *Sci. Adv.* **3**, e1700488 (2017).
22. Baran, D. *et al.* Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci.* **114**, 10900–10905 (2017).
23. Fischman, S. & Ofran, Y. Computational design of antibodies. *Curr. Opin. Struct. Biol.* **51**, 156–162 (2018).
24. Wolf Pérez, A.-M., Lorenzen, N., Vendruscolo, M. & Sormanni, P. Assessment of Therapeutic Antibody Developability by Combinations of In Vitro and In Silico Methods. in *Therapeutic Antibodies* (ed. Houen, G.) vol. 2313 57–113 (Springer US, 2022).
25. Fernández-Quintero, M. L. *et al.* Assessing developability early in the discovery process for novel biologics. *mAbs* **15**, 2171248 (2023).
26. Svilenov, H. L., Arosio, P., Menzen, T., Tessier, P. & Sormanni, P. Approaches to expand the conventional toolbox for discovery and selection of antibodies with drug-like physicochemical properties. *mAbs* **15**, 2164459 (2023).
27. Gentiluomo, L. *et al.* Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization. *Mol. Pharm.* **17**, 426–440 (2020).
28. Boughter, C. T. *et al.* Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. *eLife* **9**, e61393 (2020).
29. Akbar, R. *et al.* Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs* **14**, 2008790 (2022).
30. Khetan, R. *et al.* Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *mAbs* **14**, 2020082 (2022).
31. Raybould, M. I. J. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* **33**, 201810576 (2019).
32. Zhang, Y. *et al.* Physicochemical rules for identifying monoclonal antibodies with drug-like specificity. *Mol. Pharm.* acs.molpharmaceut.0c00257 (2020) doi:10.1021/acs.molpharmaceut.0c00257.

33. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* pro.4205 (2021) doi:10.1002/pro.4205.
34. Honegger, A. & Plückthun, A. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *J. Mol. Biol.* **309**, 657–670 (2001).
35. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* btv552 (2015) doi:10.1093/bioinformatics/btv552.
36. Oord, A. van den, Vinyals, O. & Kavukcuoglu, K. Neural Discrete Representation Learning. *Adv Neural Inf Process Syst* (2017).
37. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J. & Tagliasacchi, M. SoundStream: An End-to-End Neural Audio Codec. *IEEEACM Trans. Audio Speech Lang. Process.* **30**, 495–507 (2022).
38. Yu, J. *et al.* Vector-quantized Image Modeling with Improved VQGAN. Preprint at http://arxiv.org/abs/2110.04627 (2022).
39. Kaiser, Ł. *et al.* Fast Decoding in Sequence Models Using Discrete Latent Variables. *Proc. 35th Int. Conf. Mach. Learn.* **80**, 2390–2399 (2018).
40. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* (2019).
41. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.* 8026–8037 (2019).
42. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
43. Prihoda, D. *et al.* BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs* **14**, 2020203 (2022).
44. Wollacott, A. M. *et al.* Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Eng. Des. Sel.* **32**, 347–354 (2019).
45. Lefranc, M.-P. & Lefranc, G. Immunoglobulins or Antibodies: IMGT® Bridging Genes, Structures and Functions. *Biomedicines* **8**, 319 (2020).
46. Marks, C., Hummer, A. M., Chin, M. & Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics* **37**, 4041–4047 (2021).
47. Schmitz, S., Soto, C., Crowe, J. E. & Meiler, J. Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires. *mAbs* **12**, 1758291 (2020).
48. Gao, S. H., Huang, K., Tu, H. & Adler, A. S. Monoclonal antibody humanness score and its applications. *BMC Biotechnol.* **13**, 55 (2013).
49. Abhinandan, K. R. & Martin, A. C. R. Analyzing the "Degree of Humanness" of Antibody Sequences. *J. Mol. Biol.* **369**, 852–862 (2007).
50. Saerens, D. *et al.* Identification of a universal VHH framework to graft non-canonical antigen-binding loops of camel single-domain antibodies. *J. Mol. Biol.* **352**, 597–607 (2005).
51. Xiang, Y. *et al.* Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2. *Science* **370**, 1479–1484 (2020).
52. Sang, Z., Xiang, Y., Bahar, I. & Shi, Y. Llamanade: An open-source computational pipeline for robust nanobody humanization. *Structure* **30**, 418-429.e3 (2022).
53. Raybould, M. I. J. *et al.* Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Res.* **48**, D383–D388 (2020).
54. Lancucki, A. *et al.* Robust Training of Vector Quantized Bottleneck Models. in *2020 International Joint Conference on Neural Networks (IJCNN)* 1–7 (IEEE, 2020). doi:10.1109/IJCNN48605.2020.9207145.

55. Vaisman-Mentesh, A., Gutierrez-Gonzalez, M., DeKosky, B. J. & Wine, Y. The Molecular Mechanisms That Underlie the Immune Biology of Anti-drug Antibody Formation Following Treatment With Monoclonal Antibodies. *Front. Immunol.* **11**, 1951 (2020).
56. Vincke, C. *et al.* General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold. *J. Biol. Chem.* **284**, 3273–3284 (2009).
57. Riechmann, L., Clark, M., Waldmann, H. & Winter, G. Reshaping human antibodies for therapy. *Nature* **332**, 323–327 (1988).
58. Moutel, S. *et al.* NaLi-H1: A universal synthetic library of humanized nanobodies providing highly functional antibodies and intrabodies. *eLife* **5**, e16228 (2016).
59. Li, X. *et al.* Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies. *PLOS ONE* **11**, e0161801 (2016).
60. Abanades, B. *et al. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins*. http://biorxiv.org/lookup/doi/10.1101/2022.11.04.514231 (2022) doi:10.1101/2022.11.04.514231.
61. Cohen, T., Halfon, M. & Schneidman-Duhovny, D. NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning. *Front. Immunol.* **13**, 958584 (2022).
62. Clavero-Álvarez, A., Di Mambro, T., Perez-Gaviro, S., Magnani, M. & Bruscolini, P. Humanization of Antibodies using a Statistical Inference Approach. *Sci. Rep.* **8**, 14820 (2018).
63. Jiang, J. *et al.* Preclinical safety profile of disitamab vedotin : a novel anti-HER2 antibody conjugated with MMAE. *Toxicol. Lett.* **324**, 30–37 (2020).
64. Deeks, E. D. Disitamab Vedotin: First Approval. *Drugs* **81**, 1929–1935 (2021).
65. Aprile, F. A. *et al.* Rational design of a conformation-specific antibody for the quantification of Aβ oligomers. *Proc. Natl. Acad. Sci.* **117**, 13509–13518 (2020).
66. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).
67. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
68. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-021-01146-5.
69. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
70. Makowski, E. K. *et al.* Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. Commun.* **13**, 3788 (2022).