nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00778-3

Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV

Received: 30 March 2023

Accepted: 4 December 2023

Published online: 15 January 2024

Check for updates

Aubin Ramon[1], Montader Ali[1], Misha Atkinson[1], Alessio Saturnino[1,2], Kieran Didi[1,3,6], Cristina Visentin[4], Stefano Ricagno[4,5], Xing Xu[1], Matthew Greenig[1] & Pietro Sormanni[1]

Monoclonal antibodies have emerged as key therapeutics. In particular, nanobodies, small, single-domain antibodies that are naturally expressed in camelids, are rapidly gaining momentum following the approval of the first nanobody drug in 2019. Nonetheless, the development of these biologics as therapeutics remains a challenge. 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 reactivity with self-antigens and low toxicity. Here we present AbNatiV, a deep learning tool for assessing the nativeness of antibodies and nanobodies, that is, their likelihood of belonging to the distribution of immune-system-derived human antibodies or camelid nanobodies. AbNatiV is a multipurpose tool that accurately predicts the nativeness of Fv sequences from any source, including synthetic libraries and computational design. It provides an interpretable score that predicts the likelihood of immunogenicity, and a residue-level profile that can guide the engineering of antibodies and nanobodies indistinguishable from immune-system-derived ones. We further introduce an automated humanization pipeline, which we applied to two nanobodies. Laboratory experiments show that AbNatiV-humanized nanobodies retain binding and stability at par or better than their wild type, unlike nanobodies that are humanized using conventional structural and residue-frequency analysis. We make AbNatiV available as downloadable software and as a webserver.

Antibodies are a class of biomolecules with a 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 used as both diagnostic[2] and therapeutic agents[3]. Nanobodies (Nb) 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

[1]Centre for Misfolding Diseases, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, UK. [2]Department of Biology and Biotechnology Lazzaro Spallanzani, University of Pavia, Pavia, Italy. [3]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [4]Department of Biosciences, University of Milan, Milan, Italy. [5]Institute of Molecular and Translational Cardiology, IRCCS Policlinico San Donato, Milan, Italy. [6]Present address: Faculty of Mathematics, University of Cambridge, Cambridge, UK. e-mail: ps589@cam.ac.uk

bind to poorly accessible epitopes, and affinity and specificity at par to those of full-length antibodies[5]. Furthermore, their potential as therapeutics has gained increased recognition since the approval of the first nanobody drug, Caplacizumab, in 2019 (ref. 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 immunization[7], and second-generation in vitro techniques, relying on laboratory library construction and screening[8,9]. More recently, a third generation of approaches based on computational design has started to emerge[9]. Since the mid 1990s, in vitro methods such as phage display from naïve or synthetic libraries showed promise to replace animal immunization or other in vivo techniques to isolate new antibodies. In vitro selection is faster and cheaper than in vivo counterparts, has fewer ethical implications and enables a better control over antigen presentation[8,10]. However, despite the added costs and complexity, an increasing number of pharmaceutical and biotechnology companies prefer to obtain new antibodies by immunizing transgenic animals with a humanized immune system[11,12] or by isolating them directly from patients[13,14]. The reason for this choice is that, compared with in vitro directed evolution, antibody selection carried out by immune systems usually yields antibodies with higher developability potential and especially better in vivo properties, including long half-life, low immunogenicity, no toxicity and low cross-reactivity against self-antigens[15,16]. Up to now, most therapeutic antibodies continue to come from animal immunization[17]. 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[18–22], which remains extremely laborious with laboratory-based approaches, and in the prediction and design of biophysical properties that underpin developability[23]. Overall, computational design promises a cheaper and faster route for the discovery and optimization of antibodies, while in principle affording much better control than in vivo and in vitro techniques over other key biophysical properties such as stability and solubility[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[23–25]. However, these assays typically correlate poorly with each other, and have only been shown to correlate with selected in vivo properties in limited specific examples[16,23,26]. While advances have been made in the computational predictions of the outcome of some of these assays[27–29], or even in the number of such assays in which a lead antibody candidate is likely to perform poorly[30,31], it is 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 new deep learning method to bypass these challenges by enabling the computational engineering of antibody and nanobody 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 native variable domain (Fv) sequences 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 on the Fv sequences of human heavy chains (VH), kappa light chains (Vκ), lambda light chains (Vλ) and camelid heavy-chain single-domains (VHH).

AbNatiV can assess separately 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 of the Fv sequence, which can guide engineering by highlighting sequence regions harbouring liabilities. Therefore, AbNatiV can be useful for computational antibody design, but also to rank Fv sequences of any origin, including from in vitro discovery. The accuracy of AbNatiV in evaluating humanness is demonstrated in several benchmarks. In particular, we show that AbNatiV outperforms alternative methods when classifying antibody therapeutics. Moreover, we find that AbNatiV learns a representation of natural antibodies that captures high-order relationships between positions, which we show to be valuable for CDR grafting. We further introduce an automated humanization pipeline of antibodies and nanobodies that rely on AbNatiV. For nanobodies, this approach monitors concurrently the humanness and the VHH nativeness of a sequence. Laboratory experiments on two nanobodies binding to distinct targets show that AbNatiV-humanized nanobodies retain binding and stability at par or better than their wild type (WT), unlike nanobodies humanized with conventional structural and residue-frequency analysis.

Taken together, our results highlight the potential of AbNatiV in advancing antibody and nanobody engineering, serving as a valuable tool for computational design and ranking of Fv sequences from diverse sources, including in vitro discovery and synthetic libraries.

Results

The AbNatiV model

AbNatiV is a deep learning model trained on immune-system-derived antibody sequences. It uses an architecture inspired by that of the VQ-VAE, originally proposed for image processing (that is, for tensors of rank 3)[32]. 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 before reconstruction with a decoder (Fig. 1a). This vector quantization 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[32]. Our model contains both patch convolutional layers and transformers in the encoder and decoder (Fig. 1b). These are more suitable to capture local interactions along the sequence (that is, local motifs), and long-range interactions between such local motifs or individual residues, respectively, which may be mediated by tertiary contacts. High codebook usage (that is, high perplexity) is ensured in the bottleneck by a *k*-means initialization of the codebook and a cosine similarity search during the nearest neighbour lookup quantization, as it is needed to prevent poor data representation and maintain robust training[33] (Methods and Supplementary Fig. 1).

The model is trained with masked unsupervised learning. Unsupervised learning works on the assumption that every 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

https://doi.org/10.1038/s42256-023-00778-3

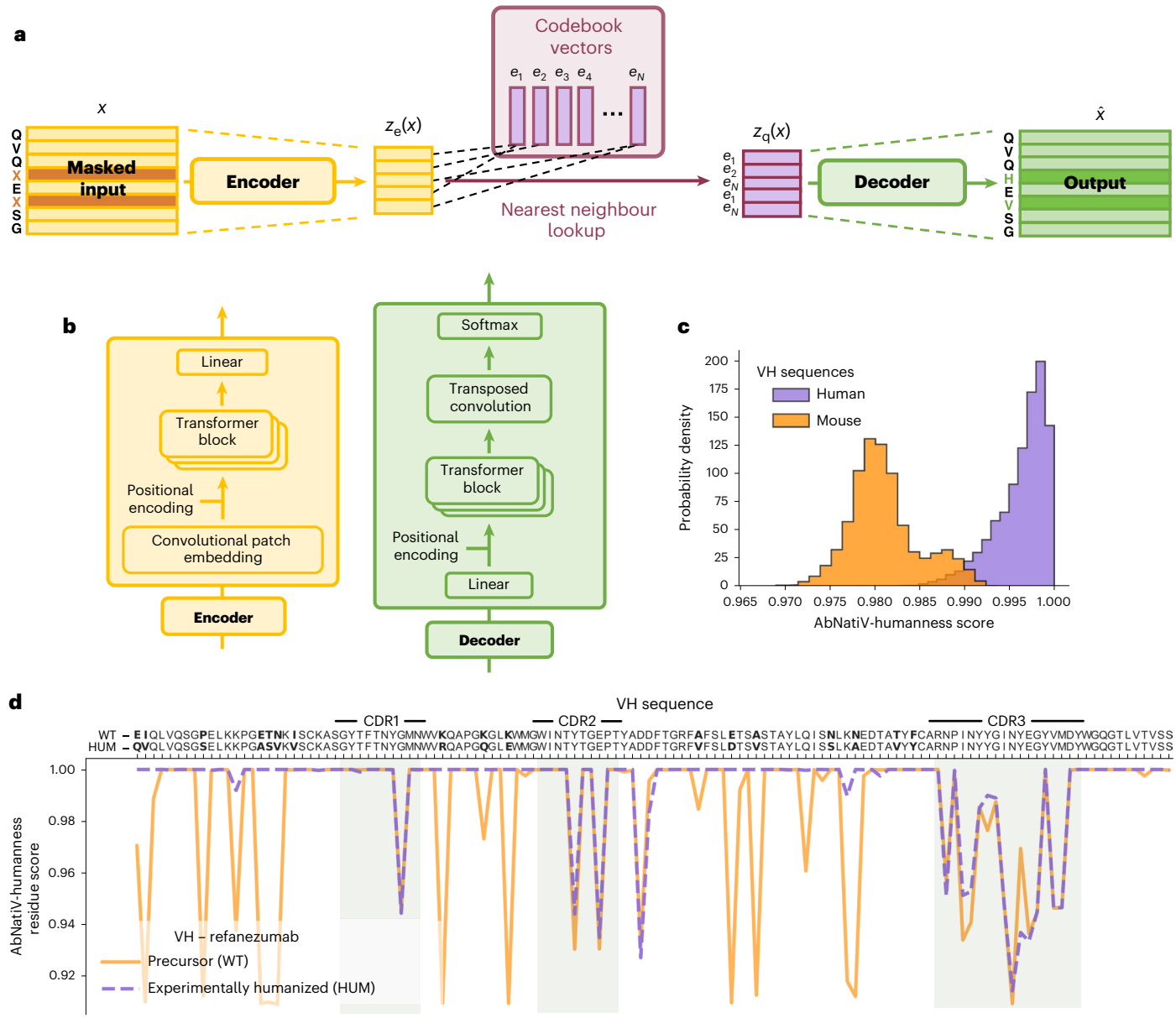


Fig. 1 | The AbNatiV model. a, Architecture of the VQ-VAE-based AbNatiV model. The one-hot encoded input sequence x is encoded into a compressed representation $z_e(x)$ through an encoder (in yellow). In the latent space (in burgundy), $z_e(x)$ is discretized with a nearest neighbour lookup on a codebook $\{e_k\}_{k=1}^{N}$ 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 a 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 humanized sequence of the refanezumab antibody therapeutic (the corresponding light chain profile is in Supplementary Fig. 5).

structure exists in the input, as is the case for natural 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 used 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 (Methods). This masking procedure is akin to a noising technique used in denoising auto-encoders[34]. From a theoretical standpoint, the approach is motivated by a manifold learning perspective, which assumes that the input data exist on a low-dimensional manifold embedded in the input space. The noising process, that is the masking and/or replacement of individual residues during training, shifts each training sequence away from the manifold of native antibodies, and the network is tasked with moving the data back onto the manifold via the output reconstruction of the input sequence. Additionally, the fact that the reconstruction loss also accounts for unmasked regions of the training sequences ensures that the network does not move data away from the manifold. Reconstruction accuracy is quantified with a mean-squared error (m.s.e.) calculated between one-hot encoded input sequences and reconstructed output sequences. Then, at inference time, the network reconstruction of

unmasked sequences represents a transformation of the input that produces an output sequence that lies closer to the manifold on which native antibodies exist. This fact establishes a crucial link between the m.s.e. of the reconstruction and antibody nativeness, as the m.s.e. can be interpreted as the distance of the input sequence from the manifold of native antibodies (Methods). Reconstruction through the network always introduces some deterioration of the perfect one-hot encoded vectors, meaning that the m.s.e. is never exactly zero, even when no residue is substituted during inference.

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 Observed Antibody Space (OAS) database[35] and other sources (Methods). The model is trained for ten epochs separately on human VH, Vκ, Vλ and camelid VHH sequences (roughly 2 million unique sequences in each training set). The κ and λ light chains are treated separately due to their substantial differences. AbNatiV takes around 1 hour per epoch to train 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 (Supplementary Fig. 2) and is used for hyperparameter optimization. Ten thousand further unique sequences, distinct from those in training and validation sets, are kept aside for testing. We observe a near-perfect overlap between the distributions of the AbNatiV scores of the training and test datasets, which supports the lack of overfitting (Supplementary Fig. 3). We further verified that there is no correlation between the AbNatiV scores of the test sequences and their median or minimum sequence difference to the training sequences ($R^2 \leq 0.002$; Supplementary Fig. 4).

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 learnt 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 0.8 represents the threshold that best separates native and non-native sequences (Methods). In the case of AbNatiV trained on VH, Vλ and Vκ human chains, this score is referred as to the AbNatiV-humanness score (Fig. 1c). Similarly, for AbNatiV trained on VHH camelid sequences, this score is referred to as the AbNatiV-VHH-nativeness score.

The sequence profile consists of one number per residue 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 humanized 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 humanization (Fig. 1d and Supplementary Fig. 5). This example shows that sequence profiles can be powerful tools to guide antibody engineering by facilitating the design of mutations to improve antibody nativeness.

Overall, AbNatiV predictions are highly interpretable, as nativeness scores tend to 1.0 with a 0.8 threshold that separates native and non-native sequences, and the sequence profile provides single-residue resolution on the sequence determinants of nativeness.

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, Extended Data Fig. 1 and Supplementary Fig. 6). For example, AbNatiV can accurately distinguish the VH human sequences of its test set from VH mouse sequences on the basis of their humanness score distribution with a PR-AUC of 0.996 (Fig. 2b) and ROC-AUC of 0.995 (Supplementary Fig. 6a). 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 a PR-AUC of 0.965 (Fig. 2b) and ROC-AUC of 0.958 (Supplementary Fig. 6a).

We further used 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 generalize to unseen sequence space. For the latter, we compiled a dataset of highly diverse human Fv sequences that we named diverse greater than 5% (at least 5% away from any sequence in the training set; Methods). As expected, classification performances on the diverse dataset slightly decrease, but overall remain very high. For the VH model, the biggest drop is found with rhesus sequences from a PR-AUC of 0.965 with the test set down to 0.923 with the diverse greater than 5% set (Fig. 2b,c). However, the VH model is still able to classify most of the diverse greater than 5% sequences as human. Only 5.5% of these sequences have a score below the nativeness threshold of 0.8, compared with 1.9% for the test VH sequences. For the light-chain models, the performances are even more comparable (Extended Data Fig. 1 and Supplementary Fig. 6), perhaps because the diverse greater than 2.5% set is less distant from the training set since diversity is more limited in light chains than in heavy chains. This performance on the control dataset is in line with our assessment of lack of overfitting (Supplementary Fig. 2), and it makes us confident in the ability of the model to generalize to sequences distant from those it was trained on.

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 (Methods and Supplementary Fig. 7). We call these datasets position-specific scoring matrix (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 greater than 0 and following the observed residue-frequency distribution; 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. We find that AbNatiV can perfectly separate real VH human sequences from PSSM-generated ones (PR-AUC of 1.000 and 0.998, respectively, for the VH human test and diverse greater than 5% datasets; Fig. 2), and that the separation is also excellent for Vκ (PR-AUC of 0.992 and 0.988, respectively; Supplementary Fig. 6a–c) and Vλ (PR-AUC of 0.990 and 0.980, respectively; Supplementary Fig. 6d–f). 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.

We then compared the performances of AbNatiV with those of other computational methods developed for the humanization of antibody sequences (Table 1, Extended Data Tables 1 and 2, Supplementary Tables 1–3 and Supplementary Figs. 8 and 9). More specifically, we focus on the recently introduced OASis 9-mer peptide similarity score[36], the Sapiens transformer model[36] and the long short-term memory network AbLSTM model[37], as these approaches were shown

https://doi.org/10.1038/s42256-023-00778-3

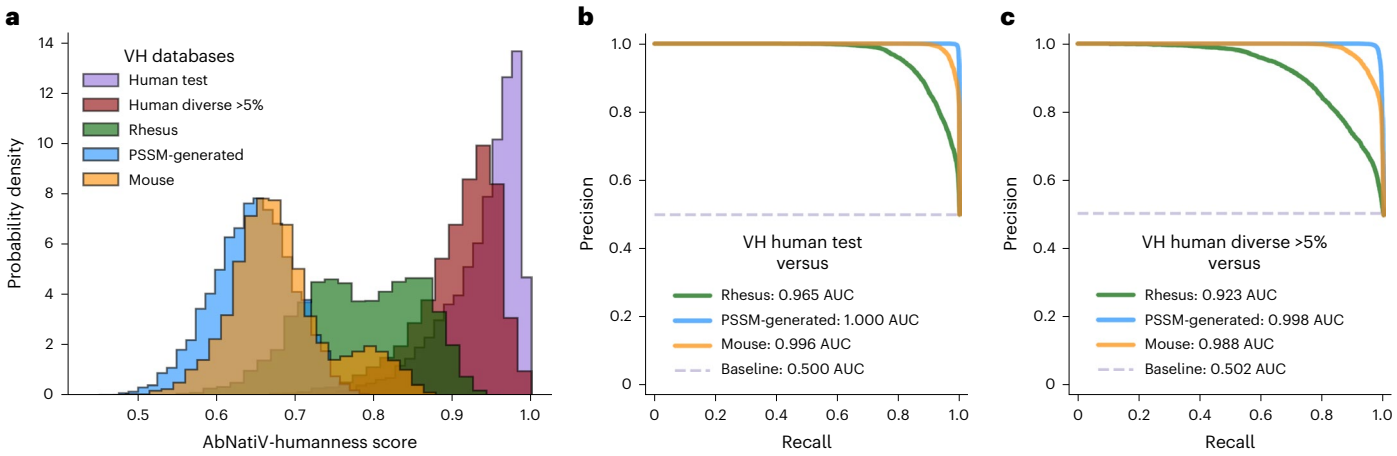


Fig. 2 | Performance on VH sequence classification. a, The AbNatiV-humanness score distributions of the human test (purple), human diverse greater than 5% (red), rhesus (green), PSSM-generated (blue) and mouse (orange) VH antibody datasets. The PSSM-generated database is made of artificial sequences randomly generated using residue positional frequencies from the PSSM of human VH sequences. The human diverse >5% dataset is made of VH sequences at least 5% different from their closest sequence in the VH training set (Methods). **b**,**c**, Plots of the PR curves of the ability of AbNatiV to distinguish the VH human test set (**b**) or human diverse >5% set (**c**) from the other datasets (see legend, which also reports the AUC). The baseline (dashed line) corresponds to the performance of a random classifier. The corresponding ROC curves are given in Supplementary Fig. 6a,b.

Table 1 | Evaluation of the PR classification and reconstruction tasks for human VH sequences

VH	Classification (PR-AUC)						Reconstruction accuracy	
	Rhesus versus		Mouse versus		PSSM-generated versus		T	D
	T	D	T	D	T	D		
AbNatiV	0.965	0.923	0.996	0.988	1.000	0.998	0.960	0.935
OASis (relaxed)	0.570	0.829	0.897	0.965	0.982	0.992	N/A	N/A
Sapiens	0.626	0.883	0.982	0.994	0.993	0.997	0.918	0.949
AbLSTM	0.721	0.892	0.963	0.986	0.998	0.998	0.807	0.856
AbLSTM retrained	0.777	0.866	0.967	0.979	0.997	0.996	0.822	0.849

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 (Methods). The first six columns report the area under the PR curve (shown in Fig. 2 and Supplementary Fig. 8), assessing the ability of the models to separate sequences in the human test (T) or the human diverse >5% (D) sets from those from mouse, rhesus and PSSM-generated (column headers). The human diverse >5% dataset is used here as a control to specifically assess the ability of the AbNatiV to generalize to sequences distant from those in its training set. 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 (N/A, not applicable). Many sequences of the D datasets belong to the Sapiens training set. Corresponding ROC results are in Supplementary Table 1.

to outperform older methods. Our results show that AbNatiV outperforms all alternative approaches on all classification tasks overall (Table 1 and Supplementary Table 1). The biggest difference is observed in the human test versus rhesus classification, where for VH sequences the AbNatiV PR-AUC is 0.965, whereas that of the best alternative method, AbLSTM, is 0.721, which increases to 0.777 once the AbLSTM architecture is retrained on our training set (Table 1). Lower performances of the alternative models are also observed for the human test versus mouse and versus PSSM-generated classification tasks. We have not included in this benchmark the recently introduced Hu-mAb method[38], since 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 of distinguishing between human and mouse sequences, we would expect it to do extremely well at the mouse versus human classification task and perhaps not as well on other tasks.

We further carried out the same benchmarks by replacing the human test set with the human diverse greater than 5% dataset, which contains sequences that are at least 5% different from any sequence in our training set. AbNatiV remains the best performing model overall. However, Sapiens marginally outperforms AbNatiV in one task: the classification of mouse sequences (by 0.006 in PR-AUC; Table 1). This result is hardly surprising, as the human diverse greater than 5% databases were built using sequences from the training sets of Sapiens and OASis[36], 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 for AbNatiV, Sapiens and AbLSTM, respectively). Sapiens reconstructs the VH sequences in the human diverse dataset slightly better than AbNatiV (94 and 95%, respectively). However, it should be noted again that the human diverse greater than 5% dataset is contained in the training set of Sapiens[36].

Similar results are found for Vκ and Vλ light chains, when comparing AbNatiV with the OASis and Sapiens methods (Extended Data Tables 1 and 2 and Supplementary Tables 2 and 3), while the AbLSTM humanness score is not defined for light chains[37]. AbNatiV exhibits higher reconstruction accuracy than Sapiens also for the light chains variable domains (VL) in the human diverse greater than 2.5% datasets (98 versus 94% for Vκ and 98 versus 93% for Vλ, respectively).

https://doi.org/10.1038/s42256-023-00778-3

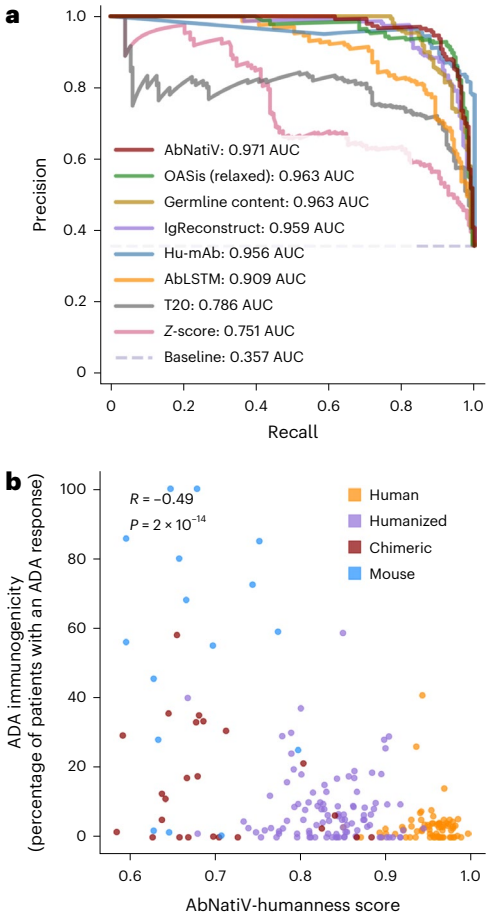


Fig. 3 | Performance on antibody therapeutics. a, Plot of the PR curves of the classification of 196 human-derived therapeutics from 353 therapeutics of non-human origin (mouse, chimeric and humanized) carried out with AbNatiV (in red) and seven other computational methods (see legend, which also reports the AUC values). The baseline (dashed line) corresponds to the performance expected from a random classifier. Corresponding ROC curves can be found in Supplementary Fig. 10. **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 two-sided P value are reported on top left corner. Sequences are coloured on the basis of their origin (that is, human in orange, humanized in purple, chimeric in red and mouse in blue).

Taken together, these results demonstrate that AbNatiV is a precise humanness assessment method that has learnt high-order relationships between residues to identify antibody sequences derived from human immune systems.

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 (that is, trained either on VH, Vκ or Vλ; 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 humanized). The PR curve (Fig. 3a) and ROC curve (Supplementary Fig. 10) are computed for AbNatiV and seven other computational approaches (Methods and Extended Data Table 3). AbNatiV outperforms all other methods when considering both AUCs with a PR-AUC of 0.971 and a ROC-AUC of 0.979. The second-best methods after AbNatiV are OASis with a PR-AUC of 0.963 and a ROC-AUC of 0.975 and Hu-mAb with a ROC-AUC of 0.979 and a PR-AUC of 0.956.

A central interest in humanization 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 who develop anti-drug antibodies (ADAs) in response to the administration of therapeutic antibodies[39]. We find that the AbNatiV-humanness score (that is, the average of the AbNatiV-humanness scores of the VH and VL; Methods) shows a Pearson correlation coefficient (R) of −0.49 ($P \cong 2 \times 10^{-14}$) with the percentage of patients who developed ADAs on 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 percentage of patients who developed an ADA response is determined in different studies carried out in drastically different ways. In particular, the dosage of the therapeutic antibody candidate and the length of the study (that is, 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 that 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 a 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.

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 single-domain antibody, and hence the ability of a VH sequence to fold independently of a VL counterpart.

We find that AbNatiV accurately discriminates VHH test sequences from the VH sequences of human (0.983 PR-AUC), mouse (0.995) and rhesus (0.992) (Fig. 4a–c and Supplementary Fig. 11). The PR-AUC between PSSM-generated artificial VHH sequences and real camelid VHH sequences from the test set is 0.942. The VHH model can classify most of the diverse greater than 5% VHH sequences as native, with a performance at par to that observed on the test set. Also, 10.4% of diverse greater than 5% VHH sequences have a score below the nativeness characteristic threshold of 0.8, compared with 10.8% for the test VH sequences. 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 (Methods). We find that AbNatiV shows higher classification performance than the retrained AbLSTM model on all tasks, and especially on the classifications with the VHH diverse greater than 5% dataset (Extended Data Table 4, Supplementary Table 4 and Supplementary Fig. 12).

CDR nativeness for grafting experiment. The grafting of target specific CDRs onto a different framework scaffold is a common technique to design an antibody with enhanced properties (for example, lower immunogenicity, higher stability or expressibility and so on)[40–42]. In the case of nanobodies, a specific camelid framework, referred to as universal framework (UF), was shown to retain very high conformational stability and prokaryotic expressibility almost independently of its CDR loops[43]. In that study, all three CDRs of six unrelated nanobodies targeting different antigens were grafted onto the UF. Binding affinity (K_D) and conformational stability (ΔG) were experimentally measured for all six

https://doi.org/10.1038/s42256-023-00778-3

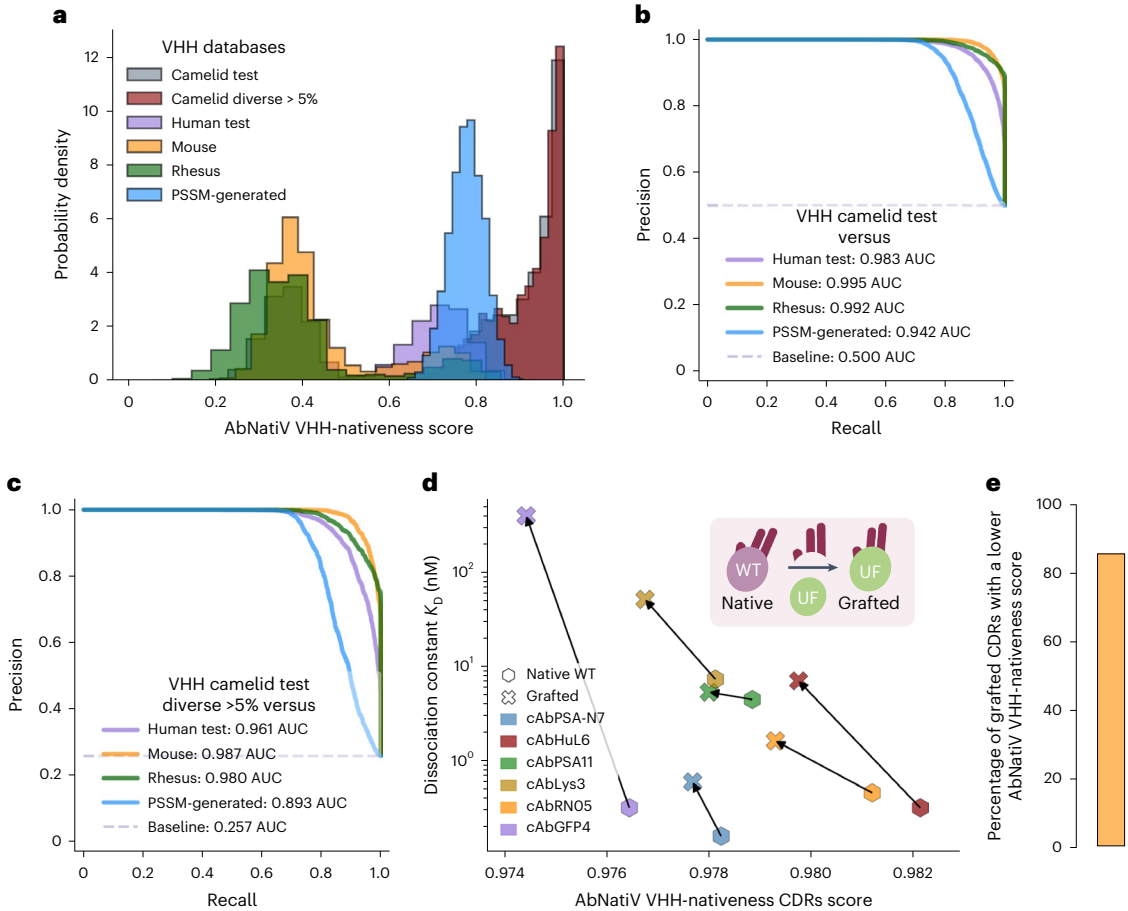


Fig. 4 | Performance on VHH sequences derived from camelids. a, The AbNatiV-VHH-nativeness score distributions of the VHH camelid test (in grey), camelid diverse greater than 5% (in red), VH human test (in purple), VH mouse (in orange), VH rhesus (in green) and VHH PSSM-generated (in blue) datasets. The VHH PSSM-generated database is made of artificial sequences randomly generated using residue positional frequencies from the PSSM of VHH sequences. The camelid diverse greater than 5% dataset is made of VHH sequences at least 5% different from their respective closest sequence in the VHH training set (Methods). Each dataset contains 10,000 sequences except the camelid diverse greater than 5%, which contains 3,468 sequences. **b**,**c**, Plots of the PR curves used to quantify the ability of AbNatiV to distinguish the VHH camelid test (**b**) or camelid diverse >5% (**c**) set from the other datasets (see legend, which also reports the AUC values). The baseline (dashed line) corresponds to the performance of a random classifier. The corresponding ROC curves are given in Supplementary Fig. 11. **d**, Plot of the binding K_D, as reported in ref. 40, as a function of the AbNatiV-VHH-nativeness score computed across all CDR positions of six nanobodies (legend) before and after grafting of all three CDRs onto a camelid UF. An arrow is directed from the native sequence in the WT framework to the grafted one. **e**, All three CDRs from a test set of 5,000 VHH sequences are computationally grafted onto the UF (Methods). The bar plot shows that 86% of them have a lower AbNatiV-VHH-nativeness score when grafted onto the UF than when they are within their native framework.

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 on grafting because of the superior stability of the UF[43]. 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 (Methods). We find that for all these six grafting examples, AbNatiV scoring anticorrelates with the experimentally measured change in binding K_D (Fig. 4d). Specifically, AbNatiV attributes a worse (lower) VHH-nativeness score to these 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 Supplementary Fig. 13.

Encouraged by these findings on six experimentally characterized 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 86% 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. 4e). Taken together, the results of these analyses indicate that AbNatiV can accurately determine whether CDR loops are in the right context.

Humanization of nanobodies

With the recent surge of interest in the use of nanobodies as therapeutics, the humanization of nanobodies has emerged as a crucial requirement to improve their therapeutic index and reduce immunogenicity risks for clinical applications[41,44,45]. Extended Data Fig. 2 depicts the AbNatiV evaluation of the humanness and VHH nativeness of three nanobody therapeutics, and of eight WT nanobodies from a SARS-CoV-2 study[46] and their humanized counterpart characterized in a separate study[44]. In that study, Sang et al. introduced a computational pipeline named Llamanade[44], which integrates structural information and residue-frequency statistics to humanize nanobody sequences. We find that all humanized nanobody sequences are assigned an AbNatiV-humanness score higher than that of their WT counterpart. This improvement of humanness affects their VHH nativeness only weakly or even improves it (Extended Data Fig. 2), which is in line with

the non-significant or very small change observed experimentally by Sang et al.[44] in the binding K_D of these nanobodies on humanization.

Encouraged by these observations, we sought to develop a framework to exploit AbNatiV for the rational humanization of nanobody sequences. By combining the humanness (VH-AbNatiV) with the VHH-nativeness (VHH-AbNatiV) assessments of AbNatiV, we propose a dual-control humanization strategy of nanobody sequences. As illustrated in Supplementary Fig. 14, this strategy begins by identifying liable positions with a low AbNatiV humanness or VHH nativeness in the residue profile. Then, it suggests potentially humanizing mutations derived from the human VH PSSM (Supplementary Fig. 7a). Finally, it accepts mutations that improve the AbNatiV-humanness score while preserving or further improving the AbNatiV-VHH-nativeness score (see Methods for further details).

Two distinct strategies to sample mutational variants are proposed, which we designate as 'enhanced' and 'exhaustive' sampling. The enhanced approach iteratively explores the mutational space, aiming for rapid convergence to identify a promising mutant. By contrast, the exhaustive approach assesses all mutation combinations within the available mutational space and selects the best sequence. It is important to note that the exhaustive sampling is considerably more computationally demanding. For instance, in the case of a sequence with ten liable positions where four mutations are allowed at each position, the mutational space encompasses 4^{10} mutants, exceeding 1 million combinations. On the other end, the enhanced sampling will explore on average less than 100 combinations of mutations. Therefore, to manage the computational complexity of the exhaustive approach, we restrict its mutational space by constraining the allowed mutations to residues enriched in both the human VH and VHH PSSMs. Conversely, the enhanced method's mutational space is larger as it restricts its allowed mutations to the human VH PSSM only. To minimize the chances of affecting antigen binding, both strategies are limited to the framework regions. For each sampling strategy, we implement both a purely sequence-based and a structure-based approach that models the nanobody structure from the input sequence (Methods). In the latter, buried residues that are not on the nanobody surface are excluded from the list of potential targets for mutations, as is commonly done in humanization strategies based on framework resurfacing[47,48].

To test the effectiveness of these different humanization pipelines we generated in silico humanized variants of two nanobodies, which we then produced and characterized in vitro. These two nanobodies bind to two distinct proteins of therapeutic relevance: Nb24 targets the β_2-microglobulin[49], and mNb6 targets the receptor-binding domain (RBD) of the Spike protein of SARS-CoV-2 (matured version of Nb6 in ref. 50). Nb24 was obtained from a llama immunization campaign and exhibits moderate binding with a dissociation constant K_D in the mid-nanomolar range[51], while mNb6 was obtained from the screening of a synthetic library and then highly optimized via saturation mutagenesis to reach a high picomolar-range K_D (ref. 52). For each WT sequence, we generated four humanized variants using the AbNatiV automated pipelines and a further control variant. Two variants were generated by each sampling method: one limited to solvent-accessible framework sites, and the other encompassing all framework sites. While the crystal structures of Nb24 and Nb6 are solved experimentally (Protein Data Bank IDs 4kdt and 7kkk, respectively), solvent-exposed sites were identified by modelling in silico the structures of the WT sequences with Nanobuilder2 (ref. 50) to simulate a more general setting in which crystal structures may not be available.

For comparison, we also generated one additional humanized variant for each WT nanobody using the automated humanization tool Llamanade that proposes humanizing mutations on the basis of structural and residue-frequency analysis[44]. We refer to these as frequency and structure-based humanized variants. All generated sequences are presented in Extended Data Table 5, and the human VH and VHH-AbNatiV profiles in Supplementary Figs. 15 and 16, which also highlight the mutations from the WT. As expected, all humanized sequences have improved humanness and similar VHH nativeness to their WT, except for the two frequency and structure-based variants that show worsened VHH nativeness (Fig. 5a,b).

WT nanobodies and all humanized designs were then produced in *Escherichia coli* and experimentally characterized (Methods).

Bio-layer interferometry (BLI) experiments show that Nb24 WT binds β_2-microglobulin with a K_D of 79 ± 6 nM (mean ± standard deviation from three independent experiments; Fig. 5c,e and Supplementary Fig. 17), which is compatible with previously reported values[51]. AbNatiV-humanized Nb24 variants obtained from both the enhanced and the exhaustive sampling strategies bind the antigen with K_D values at par to or slightly better than that of the WT (68 ± 3 and 75 ± 5 nM, respectively; Fig. 5c,e). Conversely, humanized variants containing mutations also at buried positions showed worsened K_D values, and the Nb24 variant with the most compromised binding was that from the frequency and structure-based humanization, with a K_D in the high nanomolar range (Fig. 5c,e).

We also measured the thermal stability of all produced nanobodies (Methods). We find that all Nb24 humanized variants have increased apparent melting temperatures and temperatures of unfolding onset over those of the WT (Fig. 5g). However, this improvement is the smallest for frequency and structure-based humanization; it is more pronounced for the enhanced sampling AbNatiV humanization and even larger for the exhaustive sampling strategies (Fig. 5g,i).

In agreement with previous reports[52], we find that WT mNb6 binds SARS-CoV-2 RBD with a K_D in the high picomolar range (0.78 ± 0.04 nM). The AbNatiV-humanized mNb6 variant from the enhanced sampling strategy retains this tight K_D (K_D = 0.86 ± 0.10 nM; Fig. 5d,f). However, all other mNb6 humanized variants show a binding compromised to varying degrees. The least affected variant is the one from the AbNatiV exhaustive sampling, with a K_D of 15 ± 2 nM, followed by the two AbNatiV variants that also contain mutations at buried sites. The most affected variant is the one from the frequency and structure-based humanization, which did not yield any binding signal in the assay (Fig. 5d and Supplementary Fig. 17).

In terms of stability, the enhanced sampling variants show a slight decrease of apparent melting temperature over that of the WT, but an unaffected or marginally improved temperature of unfolding onset. Conversely, the enhanced sampling variant with mutations at buried positions and the frequency and structure-based variant had decreased thermal stability, while both exhaustive sampling variants had increased thermal stability (Fig. 5h,j).

Taken together, these results underscore the effectiveness of the AbNatiV enhanced sampling humanization pipeline to enhance in silico the humanness of nanobodies by suggesting mutations that are not detrimental to binding and stability.

Discussion

In this work, we have introduced AbNatiV, a VQ-VAE-based antibody nativeness assessment method that can evaluate the likelihood of input sequences belonging 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 humanization. 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 humanization, AbNatiV exhibits higher classification performances, while often being trained on a smaller number of sequences (roughly 2 million) for fewer epochs (ten epochs). To put

https://doi.org/10.1038/s42256-023-00778-3

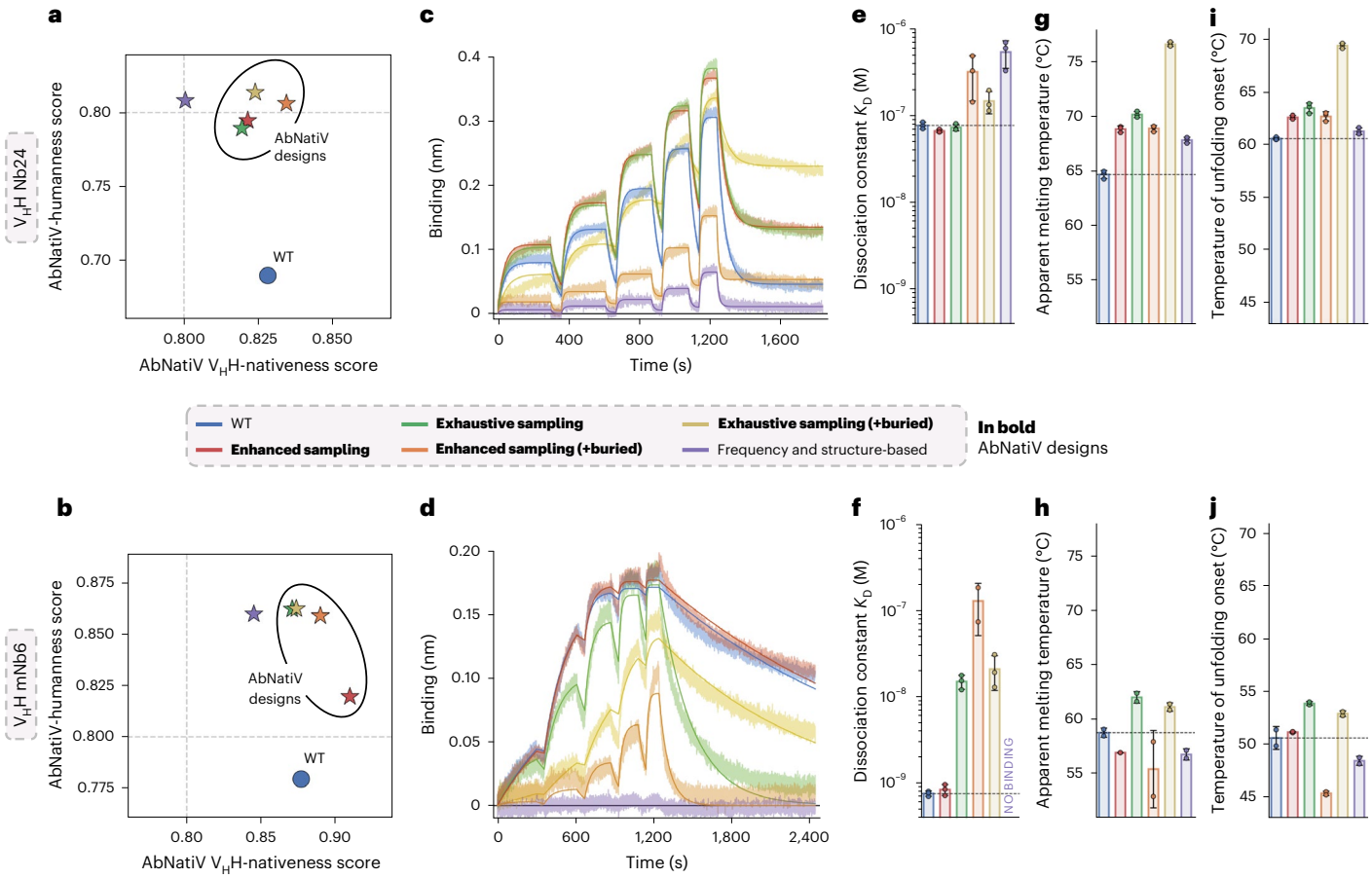


Fig. 5 | Humanization of two llama-derived nanobodies. The top row pertains to the humanization of nanobody Nb24, which binds human β_2-microglobulin, the lower row to mNb6, which binds SARS-CoV-2 RBD. In the legend, variants in bold font are different AbNatiV design strategies (text). The frequency and structure-based designs are done using the Llamanade webserver[44]. **a**,**b**, Scatter plots of the AbNatiV VH-humanness score as a function of the VHH-nativeness score for all characterized variants (legend, the WT is the blue circle): VHH Nb24 (**a**) and VHH mNb6 (**b**). **c**,**d**, BLI binding traces (associations and dissociations phases) obtained with streptavidin sensors loaded with biotinylated β_2-microglobulin (**c**) or biotinylated SARS-CoV-2 RBD (**d**). **c**, The association was monitored in wells containing 25, 50, 100, 200 and 400 nM of Nb24 nanobody variants (legend). Data were fitted globally with a 1:1 partial dissociation binding model (solid lines) using R_{max}, on rate and off rate as global parameters and $Y_{t\to inf}$ as local parameter. **d**, Association was monitored in wells containing 3.7, 11.1, 33.3, 100 and 300 nM of the WT and the enhanced sampling variants (legend); 4, 12.2, 36.4, 109.3 and 328 nM of the enhanced sampling (+buried) variant (orange) and 6.2, 18.5, 55.6, 166.7 and 500 nM of all other mNb6 variants (legend). Data were fitted globally with a 1:1 binding model (solid lines) using R_{max}, on rate and off rate as global parameters. Two additional independent BLI experiments per antigen, carried out on different days with different concentrations and times, are presented in Supplementary Fig. 17. **e**,**f**, Bar plots of the fitted K_D values from the three experiments: Nb24 (**e**) and mNb6 (**f**). **g**,**h**, Bar plots of the apparent melting temperatures: Nb24 (**g**) and mNb6 (**h**). **i**,**j**, Bar plots of the temperatures of unfolding onset (Methods): Nb24 (**i**) and mNb6 (**j**). Triplicates of the thermal stability experiments were run for the Nb24 variants, while duplicates were run for the mNb6 variants. Error bars are standard deviations.

these numbers in context, the deep VH transformer model Sapiens was trained on 20 million sequences for 700 epochs[36]. The training set size of the AbNatiV-VHH model, comprising around 2.2 million sequences, is inherently limited by the number of VHH sequences available in the literature. Conversely, for the human heavy and light chains, 2 million sequences only were used for training despite the abundance of available data for human antibody sequences. On investigation, we revealed that the VH model exhibits minimal performance improvement when expanding the training set size from 1 million to 2 million sequences (Supplementary Fig. 18a). This little gain of performance does not justify increasing the dataset training size further as this would substantially increase training time. Furthermore, having a training size comparable with that of the VHH model ensures a fair and meaningful performance evaluation across models.

AbNatiV is trained on aligned sequences. The alignment process is performed with the AHo antibody residue numbering scheme[53], which numbers each residue on the basis of its structural role (for example, being in a particular CDR loop or in the framework region). Essentially all known antibodies fit into this representation, and we posited that—albeit our method is purely sequence based—using Fv sequences aligned in this way would facilitate the learning of structural features and hence increase performance. To test this hypothesis, we used the same architecture on non-aligned sequences (Methods), which, as expected, led to a very notable performance drop. In the case of VH sequences, using non-aligned sequences resulted in a three- to fourfold decrease of both training and validation loss performances (Supplementary Fig. 18b). These findings are consistent with those of Hawkins-Hooker et al.[54], who applied a fully connected VAE to a dataset of luciferase sequences. The model trained on aligned sequences captured the information better, leading to a more successful generation of new luciferase-like sequences compared to the model trained on unaligned sequences. Moreover, using aligned sequences enables AbNatiV to produce residue profiles readily comparable across sequences of different lengths. This feature is highly advantageous for sequence engineering purposes, and for the comparison of different hits from antibody discovery or optimization campaigns.

We have also observed that AbNatiV outperforms 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 scores have a statistically significant correlation ($R = -0.5$) with the percentage of patients who developed ADA in clinical studies. This evaluation of immunogenicity with the ADA database is commonly used to benchmark immunogenicity assessment methods[36,38,55], and therefore we 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 (for example, 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 was tested only on 58 (ref. 56). In the study considered in the ADA dataset that we used, Disitamab was reported to elicit an ADA response in 58.6% of the patients. 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 four different dosages), Disitamab was shown to elicit ADA response in 23.8% of the participants[57], 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[38], showed a slightly better correlation with these ADA data ($R = -0.58$)[38]. 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 humanized 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 used.

Alongside humanness, AbNatiV quantifies the nativeness of nanobodies. The resulting model exhibits 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[41], 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 (PR-AUC of VHH 0.942, VH 1.000, Vκ 0.992 and Vλ 0.990). This observation may suggest that a bigger, and especially more diverse, VHH training dataset could be beneficial. While AbNatiV VHH is trained on slightly more sequences than AbNatiV humanness, these come from a much more restricted number of studies. Therefore, our VHH dataset has more limited diversity than the human one and it also comprises nanobodies from different camelid species (llamas, dromedaries, vicugna and so on; Supplementary Table 5), which may slightly confuse the model and demand for a larger training dataset. We expect that the publication of additional camelid immune repertoires will be beneficial for data-driven approaches such as AbNatiV, which have the potential to facilitate and accelerate nanobody development and humanization.

AbNatiV can also be used to assess whether CDR loops are in the right context or not (Fig. 4d,e). 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 on CDR grafting in a different framework (Fig. 4d). Yet, a quantitative correlation with the magnitude of the change in K_D is not observed, most probably 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 on the basis of the grafting of designed CDR loops[19,20,58]. 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 (Fig. 1d), designed CDRs or CDRs from a synthetic library.

Nanobodies exhibit substantial 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[5,44]. During the process of humanization for therapeutic purposes, it is crucial to improve humanness while preserving these traits, as they translate into high stability and binding affinity. Consequently, we introduce an automated humanization pipeline that combines the humanness and VHH-nativeness assessments of AbNatiV. We applied this dual-control strategy on two nanobodies and showed that the humanized variants generated with the enhanced sampling pipeline retain their binding activity and biophysical stability. Conversely, both properties are disrupted when conventional structural and residue-frequency humanization is applied to the same nanobodies.

We selected Nb24 and mNb6 as test nanobodies because they bind two distinct antigens with therapeutic potential, are different from each other (for example, Nb24 has a non-canonical disulfide and mNb6 has not) and represent a standard and a challenging test case, respectively, for humanization. Nb24 was obtained from immunization, and with a mid-nanomolar dissociation constant is not a particularly optimized nanobody. Conversely, with a high picomolar dissociation constant, mNb6 is a highly affinity-maturated version of a nanobody (Nb6), which was obtained from the screening of a synthetic library[52]. Consequently, one would expect that mutations in mNb6 may be more likely to disrupt affinity and stability than mutations in Nb24. Indeed, our results align with this hypothesis, with both enhanced and exhaustive sampling strategies showing excellent results on Nb24, improving both binding affinity (marginally) and stability (substantially). Conversely, only the enhanced sampling strategy did not compromise the binding of mNb6 retaining a comparable stability. In agreement with previous research[47,48], we find that framework resurfacing strategies that do not mutate buried residues are superior at preserving binding, most probably because mutations at non-solvent-exposed sites lead to slight conformational changes in the paratope region, thus affecting binding.

Overall, the enhanced sampling AbNatiV humanization yielded the most promising results. Additionally, this sampling approach is the most computationally efficient, adding to its value. Yet, the exhaustive sampling remains a valuable choice as it generates humanized sequences for different numbers of mutations via its Pareto set selection (Methods). In our experiments, we have tested only the variant with the highest VH-humanness, which is also the one with the highest number of mutations, except for the exhausted + buried strategy ran on mNb6 (Supplementary Fig. 19). Yet, this approach offers users the flexibility to pick humanized variants with fewer mutations, lowering the risk of affecting their activity or other biophysical properties.

In addition to nanobodies, AbNatiV can be used to humanize directly paired heavy and light Fv sequences by running the same sampling strategies without the VHH-nativeness constraint. In this

way, the pipeline improves both heavy- and light-chain humanness. For traditional antibodies, the whole Fv region is modelled to identify the solvent-exposed residues and residues at the VH–VL interface are not considered mutable. In this way, we limit the occurrence of mutations that could affect the pairing and relative orientation of VH and VL domain, which is important for binding.

Finally, we note that the trained AbNatiV models may facilitate applications of semisupervised learning, even if we have not explored this avenue in this work. Semisupervised learning, also known as low-N learning, combines a small amount of labelled data with a large amount of unlabelled data during training[59–61]. 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[62]. The compactness of this representation, and the fact that it was built by learning from many 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 generalize with relatively few labelled training data[60]. Approaches of semisupervised learning with protein directed-evolution data have been successfully deployed and were shown to be able to generalize to unseen regions of sequence space[59,61,63].

In summary, we expect that AbNatiV will facilitate antibody and nanobody development, as it provides a rapid, 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 new 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 such as 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.

Methods

Datasets and antibody-sequence processing

The source of all antibody sequences used for training and testing is given in Supplementary Table 5, with the full-length antibody sequences coming from the OAS[35] and the single-domain camelid VHH sequences coming from various studies[64–67]. All sequences were aligned, cleaned and processed beforehand. Non-redundant sequences were aligned using the AHo numbering scheme[53] resulting in aligned sequences of length 149. The alignment was carried out using the widely used ANARCI software[68] 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 intradomain disulfide 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. This script maximizes the identity between the multiple sequence alignment 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 did not have two cysteines at the conserved positions (because of, for example, sequencing errors) were discarded. Furthermore, Fv sequences with more than one or two missing residues at the N- and the C-terminal, respectively, 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, a leucine or a lysine, respectively, were added at the C terminus (AHo 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 (from human, rhesus and mouse) and VHH antibody sequences from various studies from the literature were assembled (Supplementary Table 5) and processed as described above. All the parsed sequence datasets used in this study are available online in the AbNatiV GitLab at https://gitlab.developers.cam.ac.uk/ch/sormanni/abnativ.

Training, validation, test and diverse datasets. A total of 2,000,000 sequences from the human heavy, lambda and kappa databases were used to train three distinct models, respectively, and 2,144,185 sequences from the VHH databases (camelid and PDB-sdAB) 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. These training, validation and test sequences were selected as random splits from the larger database of unique sequences. As we only had unique aligned Fv sequences, this procedure ensured that sequences in training, validation and test datasets were at least one mutation away, as commonly done in the field when dealing with large databases of sequences.

Furthermore, to be able to assess performance on a dataset of sequences that were more distant from any training sequence, we built an additional diverse dataset for each model. Such diverse datasets were compiled with sequences that were at least 5% different from any sequences of the training set (2.5% for Vκ and Vλ, as light chains have less diversity). Percentage difference is defined as the number of mutations between an aligned test sequence and an aligned training sequence (gap to gap is not considered a mutation), divided by the length of the gapless test sequence. For the human models (VH, Vκ and Vλ) diverse sequences were extracted from both the test and BioPhi datasets (subset of the training dataset of the Sapiens transformer from BioPhi[36]; Supplementary Fig. 20) to yield the corresponding diverse greater than 5% (or greater than 2.5% for the light chains) dataset. For the VHH model, diverse sequences were extracted from the test dataset by requiring at least 5% difference from the closest sequence in the training set. Supplementary Fig. 20 shows the cumulative distribution functions of the minimum percentage different from training sequences for each dataset. Supplementary Fig. 4 shows the distribution of the sequence difference between training sequences and all sequences in the datasets used to assess AbNatiV performance, as well as the lack of correlation between the AbNatiV nativeness score and the distance of that sequence from the training set.

PSSM-generated datasets of artificial sequences. Position weight matrices (PWM) and corresponding PSSMs were computed from each human and camelid antibody training datasets (Supplementary Fig. 7). 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; Supplementary Fig. 7) considering only those amino acids enriched at that position (that is, PSSM log-likelihood score greater than zero).

The AbNatiV model

VQ-VAE architecture. The AbNatiV model takes aligned antibody sequences of length 149 as input, and one-hot encodes each into a tensor of the dimensions 149 × 21. 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 VQ-VAE framework[32], which involves a VAE with a discretization of the dense latent space through code vectors (Fig. 1a). The sequence input $x \in \{0,1\}^{149\times 21}$ 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. To discretize $z_e(x)$ 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 quantized embedding $z_q(x) \subset \mathbb{R}^{l\times d_c}$. Finally, $z_q(x)$ is decoded to generate the reconstructed output $\hat{x} \in \{0,1\}^{149\times 21}$ having the original dimensions as the original sequence input x.

For increased codebook usage (that is, 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. 69, 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 normalized. The Euclidean distance of the l_2-normalized vectors is used during the nearest neighbour lookup resulting in a cosine similarity search as proposed in the image modelling model ViT-VQGAN[70]. Furthermore, the code vectors from the codebook are updated during training by exponential moving average with a decay of 0.9 to assure a more stable training[71].

The encoder and decoder layers are illustrated in Fig. 1b. In the encoder, the input sequence is embedded by a patch convolutional layer[70]. A one-dimensional (1D)-convolution layer with a kernel size K equal to its stride S embeds each of the non-overlapping patches of dimension $K \times 21$ into a single vector of size d_{emb} (that is, 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[72], with H heads in the multihead attentions layer and a hidden dimension d_{ff} in the feed forward layer. Before quantization, 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 hyperparameters 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 the hyperparameters were manually tuned for the VH and VHH models. It has been found empirically that the same hyperparameter values lead to the best performances for both models. Since the hyperparameters do not look to be dependent on the origin of the training set, the same hyperparameter values were used across all models and their values are given in Supplementary Table 6.

Unsupervised masked learning. Like the original VQ-VAE[32] the AbNatiV models are trained to minimize a negative evidence lower bound (NELBO) consisting of three terms as follows:

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

The first term is the negative log-likelihood reconstruction loss, which is characteristic of the VAEs. This term is approximated by the reconstruction m.s.e. between the input x and the decoder output $\hat{x}$. The second and third terms are associated with the vector quantization step in the latent space, enabling the codebook to be trained. Both terms are m.s.e.s between the encoded input $z_e(x)$ and the quantized latent embedding $z_q(x)$. In particular, the second term, stop gradient, sg, is applied to $z_e(x)$ to detach it from the computational graph, thereby 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 stop gradient allows the code vectors and the encoder to be updated at different speeds. The relative learning speed between these two terms is imposed by the scaling factor β. In all our models, β is set to 0.25. By choosing $\beta < 1$, the code vectors are updated more rapidly to align with the encoder, preventing an arbitrary growth of the encoder outputs[32].

The neural network is implemented using PyTorch v.1.14 (ref. 73) and enhanced by the PyTorchLightning.0.7 module. The models are trained with a batch size of 128 by the Adam optimizer[74] with a learning rate of 4×10^{-5}. During training, a masking is applied to the one-hot encoded inputs. As in the training of the language transformer model BERT[72], 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 one in 21 for each residue, which we use as a mask token; 10% are randomly replaced by another residue or gap and 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).

Training with non-aligned sequences

For comparison, we trained the same VQ-VAE architecture (the same hyperparameters and number of training epochs) on non-aligned VH sequences. A padding of the value zero was added to the left and right of the one-hot input vectors of non-aligned sequences to reach a size of 149. If the padding size required was odd, one more pad was added to the right side. The loss function was identical. For the reconstruction accuracy, only the non-padded components were considered.

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:

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

where $\|\hat{x}_i - x_i\|_2^2$ is the m.s.e. at sequence position i between the aligned input sequence x and the reconstructed output sequence $\hat{x}$ of a trained AbNatiV model. This m.s.e. is summed over all 149 positions of the aligned sequence and normalized by the length of the input sequence (that is, 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 m.s.e. X was 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 $(0.8-1) \times (Y-1)/(T_R - 1) + 1$ 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 used to rank different sequences (Supplementary Fig. 21). 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 is one classified as non-native. T_R is calculated for each trained model as follows. The PR curves are generated between human sequences (human test and human BioPhi datasets) as positives, and non-human sequences (mouse) as negatives for the VH, Vκ and Vλ models. Similarly, the PR curve is also calculated between VHH sequences (camelid test) as positives and non-VHH sequences (human test and house) as negatives, all computed on the $Y = \exp(-X)$ scored sequences (Supplementary Fig. 21a,d,g,j). For every model, the PR optimal threshold value T_R is extracted as the point closest to (1,1) (Supplementary Fig. 21b,e,h,k, $T_R(\text{VH}) = 0.988047$, $T_R(\text{V}\kappa) = 0.992496$, $T_R(\text{V}\lambda) = 0.985580$ and $T_R(\text{VHH}) = 0.990973$). 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 (Supplementary Fig. 21c,f,i,l). 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 T_R 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, Vκ and Vλ 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 m.s.e. reconstruction error at each position of the given sequence.

Performance metrics

All the performance metrics reported are computed by analysing 10,000 scored sequences for each database, except for the diverse datasets (Supplementary Table 5). For datasets smaller than 10,000, the whole dataset is used.

Classification. The AUC of the ROC and of the PR curves are computed to quantify the ability of a model to classify sequences. For ROC curves, the AUC is equal to one 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 one 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.

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:

$$\text{reconstruction accuracy} = \frac{\sum_{i=1}^{149} 1_{x_i = \hat{x}_i}}{149}$$

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

Benchmarking with other assessments from the literature. Open-source antibody humanness assessments from the literature were used to benchmark the performances of AbNatiV. These assessments include OASis and Sapiens from Biophi[36] and AbLSTM[37].

OASis is an average 9-mer peptide similarity searched through the OAS database. Sapiens is an unsupervised human antibody language model based on the transformer encoder BERT[72] network. It is trained on unaligned human antibody sequences from the OAS database. The GitHub implementation (https://github.com/Merck/BioPhi) of OASis and Sapiens 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.

AbLSTM[37] is an unsupervised long–short-term-memory (LSTM) neural network. Human heavy chains sequences from the OAS database are aligned before training. Here, we used the pretrained model in the benchmarking, and we also retrained the AbLSTM for ten 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 pretrained 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 2×10^{-3}). The negative log sum loss of the AbLSTM model was used as its humanness or VHH-nativeness scores as done in the original work[37].

Predictions on antibody therapeutics

Here, 549 antibody therapeutics from the IMGT database[75] were obtained from the BioPhi dataset[36]. This dataset includes 196 fully human therapeutic sequences and 353 therapeutics of non-human origin (mouse, chimeric and humanized). 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 ADA response during clinical trials, were also obtained from the BioPhi dataset[36]. 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. 36. The alternative methods considered in this work are the BioPhi germline content[36] (sequence identity to closest human germline), Hu-mAb[38] (random forest-based humanness), IgReconstruct[76] (positional nucleotide frequency scoring from back-translated human antibodies), AbLSTM[37], T20 (ref. 77) (similarity average among the closest 20 sequences) and Z-score[78] (similarity average across all sequences) assessments. Light-chain-only antibodies (that is, istiratumab, lulizumab pegol, placulumab and tibulizumab) are removed from the IMGT BioPhi parsed dataset as the original pretrained AbLSTM can only score heavy chains. Because the Fv sequence of pexeluzimab has missing C-terminal 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 Data 1 and 2.

Grafting assessment on nanobodies

In ref. 40 all three CDRs of six nanobodies were grafted onto a camelid VHH framework sequence, referred to as the UF. Binding K_D and conformational stability ΔG were experimentally measured for all six WT nanobodies, and corresponding variants with CDRs grafted onto the UF. Here, we compute the nativeness scores of the six 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 CDR scores. These are defined as the sum of the m.s.e. reconstruction scores of all residues at the CDR positions (according to the AHo numbering scheme) normalized 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 CDR 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). All the sequences with their respective K_D and AbNatiV-VHH-CDR score are available in Supplementary Data 3.

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

Humanness assessment of nanobodies

For the analysis reported in Fig. 5, 300 VH human sequences and 300 camelid sequences randomly selected from the test datasets are scored both with the AbNatiV human heavy and camelid heavy models to provide background distributions. Then, we further scored eight WT nanobodies from a SARS-CoV-2 study[46] and their humanized counterpart as reported in ref. 44, and three therapeutic nanobody sequences (envafolimab, caplacizumab and rimteravimab) available from the therapeutic database Thera-SAbDab[79].

Automated humanization of nanobodies

The humanization process of nanobody sequences by AbNatiV follows a dual-control strategy that seeks to increase the humanness while retaining the VHH nativeness of a given sequence. Standard antibodies can be humanized exactly as described here by removing all steps involving the VHH nativeness.

Given an input sequence, the VH-AbNatiV and VHH-AbNatiV residue profiles are computed along with the solvent-accessible surface area (SASA) using the 'rolling ball' algorithm[80] on the whole unbound structure modelled with NanoBuilder2 from the ImmuneBuilder software[50]. The SASA of each residue is converted into a relative SASA (RASA) value by dividing the SASA of the given residue X under scrutiny with its maximum allowed SASA[81]. The latter is obtained as the SASA of residue X in the context of the Gly-X-Gly tripeptide in a fully extended conformation. Structural modelling and SASA calculations are only performed when the user chooses to do framework resurfacing: that is, to avoid mutating any buried residue, which is the default behaviour.

To reduce the mutational space, we first flag positions for mutation using the residue nativeness profiles. The search is restricted to the framework region, as CDRs typically contain binding residues. Furthermore, if framework resurfacing is selected as an option, mutable residues must exhibit a RASA greater or equal to 15%. By comparison, in the work of Chen et al.[82] a RASA of 20% serves as a cut-off between buried and exposed residues. Starting from these automatically identified mutable positions, we developed two distinct sampling methods to explore the mutational space.

In these two pipelines, all the sampling parameters are fully adjustable (for example, tolerance of humanness, VHH-nativeness decrease or buried residues). Users can also look at the AbNatiV residue profiles and make in-depth analyses of the expected impact of humanization. This empowers users to make fully informed decisions when designing their humanized sequences and selecting those for experimental testing.

Enhanced sampling. The enhanced sampling is illustrated in Supplementary Fig. 14a. Flagged positions have a VH-AbNatiV score smaller or equal to 0.98. Convergence towards the best combination of mutations is achieved by mutating each position subsequently one at a time, as opposed to exploring all possible combinations. The order in which positions are mutated is defined starting from those mutable positions that are least affected when other positions are mutated. This strategy increases the odds that positions mutated early remain stable even after subsequent mutations along the sequence are performed, leading to a more efficient path towards identifying the best mutational variant. Thereby, a first calculation is performed to sort positions to mutate on the basis of their average interdependence on mutations at every other position in the sequence. To quantify this dependence, a computational deep mutation scanning is implemented. For a given position, each of the other positions is individually mutated into all available amino acid residues (19 possibilities). For each mutation, and each of the other positions, we calculate the difference between the AbNatiV-VHH residue score at the position under scrutiny of the WT sequence and that of the mutated sequence (note that mutations are at other positions but may still affect the score of this position and this is what we are probing for here). These differences are then averaged into a single value quantifying the dependence of the position under scrutiny on mutations elsewhere in the sequence. This procedure is iterated for every liable position.

Subsequently, starting from the position with the least dependence on mutations at other positions, we mutate it with all the amino acids significantly enriched in the human VH PSSM (that is, with a PSSM log-likelihood score greater than 0 and a PWM frequency greater than 0.01; Supplementary Fig. 7). We exclude cysteines and methionines from the list of candidate mutations as these are linked to developability liabilities. The selected mutation at each position is then the one that increases most the multi-objective function: $0.8\Delta VH + 0.2\Delta VHH$ and that does not decrease the VHH-AbNatiV score by more than 1.5% of that of the WT (that is, 1.5% decrease tolerance for ΔVHH). If no such mutation is found (for example, all screened ones decrease the VHH nativeness by more than 1.5%), the residue is left to WT and the procedure continues to the next mutable position. If a mutation is found, the sequence is updated and the process of selecting positions for mutation in Fig. 15a recommences from the beginning to ensure that no over other positions has become a liability (that is, residue score less than or equal to 0.98) following the introduction of this new mutation.

Exhaustive sampling. The exhaustive sampling is illustrated in Supplementary Fig. 14b. Flagged positions have either a VH-AbNatiV or VHH-AbNatiV score smaller than or equal to 0.98, or the WT residue is not enriched in the human VH PSSM (that is, does not have a PSSM log-likelihood score greater than 0 and a PWM frequency greater than 0.01; Supplementary Fig. 7). We generate all the possible combinations of mutations at all liable positions by considering as candidates for each position those amino acids significantly enriched in both human VH and VHH PSSMs (that is, with a PSSM log-likelihood score greater than 0 and a PWM frequency greater than 0.01; Supplementary Fig. 7). Cysteines and methionines are excluded from the list of candidates as these are linked to developability liabilities. The WT residue is retained in the list of candidate amino acids at each liable position. First, we retain only those combinations of mutations that do not decrease the VHH-nativeness score by more than 1.5% over that of the WT. Then, we compute the Pareto front that maximizes the VH-humanness score while minimizing the number of mutations over all remaining combinations of mutations. In fact, given that WT residues were retained in the list of candidate amino acid substitutions, the method produces mutational variants that have a number of mutations ranging from 0 (the WT, which is one possible combination) and the total number of identified liable positions.

At the end, this approach returns a set of mutational variants with the highest VH-humanness for each number of mutations that are beneficial to the VH-humanness (Supplementary Fig. 19). In the Pareto analysis, increasing the number of mutations is beneficial only when it further increases the VH-humanness score. For instance, we see in Supplementary Fig. 19d that going from nine to ten mutations does not increase the VH-humanness further, and therefore the variant with ten

mutations is not selected in the Pareto front. In this work, experimental testing was conducted exclusively on the sequence exhibiting the highest humanness score, which happens to be the one with the highest number of mutations in all exhaustive sampling designs, except for the variant in Supplementary Fig. 19d.

Frequency-based and structure-based nanobody humanization. To provide a benchmark for the AbNatiV humanization pipelines described above, we carried out nanobody humanization also using the recently introduced Llamanade humanization pipeline[44]. This approach builds on a systematic analysis of the sequence and structural properties that distinguish nanobodies from human VH, and proposes humanizing mutations on the basis of the analysis of the input nanobody modelled structure and the key differences between its sequence and sequences of human VH domains. These frequency- and structure-based designs were carried out with the Llamanade webserver accessed on 4 July 2023 (at http://35.208.211.136).

Protein production

Genes encoding the Nb24 and mNb6 WT nanobodies and their humanized variants were synthesized and cloned into an isopropyl-β-D-thiogalactopyranoside (IPTG)–inducible vector (by GenScript in vector pET29a(+)), including a leading PelB sequence to enable translocation to the periplasm and facilitate intradomain disulfide bond formation and ultimately the secretion of the protein to the expression media. A C-terminal 6× His tag is added for purification. All expressed amino acid sequences are given in Extended Data Table 5. Care was taken to maintain the same codon usage as the WT, except for the mutated amino acid positions. Plasmids were transformed into *E. coli* Shuffle LysY strain to further facilitate the formation of the disulfide bond, and to enable the secretion to the expression media (which is facilitated by the LysY leakier cell wall). Cultures (0.5 l) of Luria-Bertani media were inoculated at initial 0.03 OD_{600} (optical density at 600 nm), grown at 37 °C until reaching 0.8 OD_{600} and then induced with 500 µM IPTG at 30 °C for overnight expression.

His Mag Sepharose Excel magnetic beads (Cytiva) were washed in PBS and added to the cultures (1 ml per 0.5 l) about 3 h before harvesting to capture the secreted his-tagged nanobodies. Loaded beads were then fished out from the expression media using an AmMag magnetic wand (GenScript) and purification was performed with an AmMag SA Plus Semiautomated System (GenScript) using PBS as running buffer and carrying out washing steps with PBS 4 mM imidazole, and elution with PBS 200 mM imidazole. Eluted nanobodies were further purified by size exclusion chromatography using a Superdex 75 10/300 column equilibrated in PBS on an Akta Pure System (Cytiva) to remove the imidazole, further increase the purity and isolate monomeric nanobodies. Purified nanobodies were aliquoted, flash-frozen in liquid nitrogen and stored at −80 °C. Each aliquot was used only once and, following thawing, was centrifuged at 21,000*g* at 4 °C for 10 min to pellet down any precipitate that may have formed during freeze–thawing.

Recombinant β_2-microglobulin was expressed and purified to homogeneity as previously reported in ref. 83. Briefly, *E. coli* BL21(DE3) cells were transformed with pET29b carrying the coding sequence of β_2-microglobulin. The transformed cells were grown at 37 °C in Luria-Bertani medium supplemented with kanamycin and protein expression was induced with 1 mM IPTG for 3 h. β2-microglobulin was purified from the inclusion bodies. The cell pellet was resuspended in Triton buffer (100 mM sodium phosphate pH 7.4, 0.1% Triton, 1 mM EDTA, 10 mM DTT) supplemented with lysozyme and DNase. The cells were lysed by sonication and then centrifuged. The pellet obtained was washed with Triton buffer and then dissolved in 6 M GuHCl. β_2-microglobulin was refolded by consecutive dialysis (20 mM sodium phosphate pH 7.4, 150 mM NaCl; 20 mM sodium phosphate pH 7.4, 75 mM NaCl; 20 mM sodium phosphate pH 7.4, 35 mM NaCl and 20 mM Tris-HCl pH 8.3), and then purified by ion exchange using a Hi Prep Q FF 16/10 column (GE Healthcare Life Sciences) connected to an Akta Pure system (Cytiva). The protein was eluted with a linear 0–1 M NaCl gradient in 20 mM Tris-HCl pH 8.3. Purified β_2-microglobulin was aliquoted, lyophilized and stored at −80 °C. SARS-CoV-2 RBD was purchased as biotinylated purified protein from CUSABIO (product code CSB-MP3324GMY1-B) and stored at −80 °C.

Protein concentrations were measured using blanked absorbance 280 nm values and extinction coefficients calculated from the amino acid sequence using the Expasy ProtParam tool (web.expasy.org/protparam/).

LC–MS

The mass of all antibodies was verified by liquid chromatography with mass spectrometry (LC–MS) using an ACQUITY UPLC/VionTM-IMS-QTof system coupled with an electrospray ionization source. Liquid chromatographic separation of samples was performed on ACQUITY UPLC Protein BEH C4 column (300 Å pore diameter, 1.7 µm, 2.1 × 50 mm, Waters) using gradient elution. Then 1 µl of sample was injected with a flow rate of 0.3 ml min^{-1} and the analysis was carried out at default parameters. The acquired data was processed using UNIFI software. Disulfide bonds (−2 Da per bond) were detected in all variants (Extended Data Table 5).

β2-microglobulin biotinylation

To enable BLI binding assays with streptavidin sensors, β_2-microglobulin was biotinylated. Next, 10 µM of β_2-microglobulin were incubated with 1× molar concentration of EZ-Link Sulfo-NHS-LC-Biotin (Thermofisher 21335) for 2 hours, quiescent at room temperature. After this time, unreacted biotin was removed by size exclusion chromatography using a Superdex 75 10/300 column equilibrated in PBS on an Akta Pure System (Cytiva). Biotinylated β_2-microglobulin was then characterized with LC–MS to determine the degree of labelling (Supplementary Fig. 22).

Measurements of thermal stability

Measurements of apparent melting temperature were carried out in PBS at 6 µM nanobody concentration (except for mNb6 exhausted sampling + buried, which was at a concentration of 1.5 µM because of insufficient material) on a Tycho system (NanoTemper). Each experiment was repeated three times for Nb24 variants and twice for mNb6 variants. Each 350/330 fluorescence ratio trace is first smoothed via a Savitzky–Golay filter (window length 21, polynomial order two) and fitted with the two-state thermal denaturation model:

$$y = \frac{\alpha_N + \beta_N T + (\alpha_D + \beta_D T)\exp\left(\frac{\Delta H_{D-N}}{R}\left(\frac{1}{T_M} - \frac{1}{T}\right)\right)}{1 + \exp\left(\frac{\Delta H_{D-N}}{R}\left(\frac{1}{T_M} - \frac{1}{T}\right)\right)}$$

where α_N, β_N and α_D, β_D are the intercept and slope of the linear baselines of the native (N) and denatured (D) states, respectively, R is the gas constant, $\Delta H_{D\text{-}N}$ is the enthalpy of equilibrium between the native and the denatured state, and T_M is the apparent melting temperature. Each 350 to 330 nm fluorescence ratio trace is first smoothed via a Savitzky–Golay filter (window length 21, polynomial order 2) and then fitted. The temperature of unfolding onset T_{onset} is defined as the temperature needed to unfold 5% of the folded population. By definition, T_{onset} is a function of T_M and $\Delta H_{D\text{-}N}$:

$$T_{onset} = \frac{T_M}{1 - T_M \frac{R}{\Delta H_{D-N}} \ln \frac{0.05}{0.995}}$$

BLI affinity measurements

BLI measurements were performed using an OctetBLI K2 system (ForteBio). All assays were carried out in PBS supplemented with 0.05% Tween-20 (Sigma) to suppress non-specific interactions with the sensors.

All assays were carried out in a black 96-well plate (Greiner 655209), 200 µl per well, and all sensors were subjected to prehydration in the assay buffer for at least 15 min before usage. The assay plate was kept at 30 °C with a shaking speed of 1,000 r.p.m. The loading wells contained 50 nM of biotinylated β_2-microglobulin or 30 nM of biotinylated SARS-CoV-2 RBD (purchased from CUSABIO, product code CSB-MP3324GMY1-B). All experiments consisted in a baseline step, a loading step, another baseline step, followed by several association and short dissociation steps. After the last association step, a long dissociation step is performed. The number of association and/or dissociation steps, their time and analyte concentrations used varied among experiments (Fig. 5 and Supplementary Fig. 17 and their captions). In all experiments a reference sensor (loaded in the same way as the assay sensors but probing only buffer wells in all association steps) was used and its signal was subtracted from that of each assay sensor before data analysis. Binding data of all Nb24 nanobody variants were fitted globally with a 1:1 partial dissociation binding model using R_{max}, on rate and off rate as global parameters and $Y_{t\to inf}$ as local parameter. Data of all mNb6 variants were fitted globally with a standard 1:1 binding model using R_{max}, on rate and off rate as global parameters.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the large training, validation and testing datasets needed to train and evaluate AbNatiV are available online in the AbNatiV GitLab at https://gitlab.developers.cam.ac.uk/ch/sormanni/abnativ/-/tree/main/datasets?ref_type=heads (https://doi.org/10.5281/zenodo.10171047, ref. 84). Details and sources of these datasets are presented in Supplementary Table 5. Smaller datasets required to analyse the therapeutic classification, ADA correlation and VHH grafting studies are compiled in the Supplementary Data file, with details and sources included in the legends. All the sequences that were tested in vitro are provided in Extended Data Table 5.

Code availability

The AbNatiV code repository including the trained models and the automated humanization pipeline is available at https://gitlab.developers.cam.ac.uk/ch/sormanni/abnativ (https://doi.org/10.5281/zenodo.10171047, ref. 84). A user-friendly webserver to run AbNatiV is provided at www-cohsoftware.ch.cam.ac.uk/index.php/abnativ. To access the webserver, users need to register a free account and log in.

References

1. Goldman, R. D. 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. Peyvandi, F. et al. Caplacizumab for acquired thrombotic thrombocytopenic purpura. *New Engl. J. Med.* **374**, 511–522 (2016).
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.* https://doi.org/10.1039/d2cb00231k (2023).
11. Murphy, A. J. et al. Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. *Proc. Natl Acad. Sci. USA* **111**, 5153–5158 (2014).
12. Lee, E.-C. et al. Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nat. Biotechnol.* **32**, 356–363 (2014).
13. 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).
14. Wrammert, J. et al. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453**, 667–671 (2008).
15. Lonberg, N. Fully human antibodies from transgenic mouse and phage display platforms. *Current Opin. Immunol.* **20**, 450–459 (2008).
16. Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl Acad. Sci. USA* **114**, 944–949 (2017).
17. Lu, R.-M. et al. Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* **27**, 1 (2020).
18. 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).
19. Sormanni, P., Aprile, F. A., Vendruscolo, M. & Tessier, P. M. Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. *Proc. Natl Acad. Sci. USA* **112**, 9902–9907 (2015).
20. Aguilar Rangel, M. et al. Fragment-based computational design of antibodies targeting structured epitopes. *Sci. Adv.* https://doi.org/10.1126/sciadv.abp9 (2022).
21. Baran, D. et al. Principles for computational design of binding antibodies. *Proc. Natl Acad. Sci. USA* **114**, 10900–10905 (2017).
22. Fischman, S. & Ofran, Y. Computational design of antibodies. *Curr. Opin. Struct. Biol.* **51**, 156–162 (2018).
23. Wolf Pérez, A.-M., Lorenzen, N., Vendruscolo, M. & Sormanni, P. Assessment of therapeutic antibody developability by combinations of in vitro and in silico methods. *Methods Mol. Biol.* **2313**, 57–113 (2022).
24. Fernández-Quintero, M. L. et al. Assessing developability early in the discovery process for novel biologics. *mAbs* https://doi.org/10.1080/19420862.2023.2171248 (2023).
25. 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* https://doi.org/10.1080/19420862.2022.2164459 (2023).
26. Gentiluomo, L. et al. Advancing therapeutic protein discovery and development through comprehensive computational and biophysical characterization. *Mol. Pharm.* **17**, 426–440 (2020).
27. Boughter, C. T. et al. Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. *eLife* **9**, e61393 (2020).
28. Akbar, R. et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs* https://doi.org/10.1080/19420862.2021.2008790 (2022).
29. Khetan, R. et al. Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *mAbs* https://doi.org/10.1080/19420862.2021.2020082 (2022).

30. Raybould, M. I. J. et al. Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl Acad. Sci. USA* **116**, 4025–4030 (2019).
31. Zhang, Y. et al. Physicochemical rules for identifying monoclonal antibodies with drug-like specificity. *Mol. Pharm.* **17**, 2555–2569 (2020).
32. Van Den Oord, A., Vinyals, O. & Kavukcuoglu, K. Neural discrete representation learning. *Adv. Neural Inf. Process. Syst.* **2017**, 6307–6316 (2017).
33. Lancucki, A. et al. Robust training of vector quantized bottleneck models. In *Proc. 2020 International Joint Conference on Neural Networks (IJCNN)* 1–7 (IEEE, 2020); https://doi.org/10.1109/IJCNN48605.2020.9207145
34. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proc. 25th International Conference on Machine Learning* 1096–1103 (Association for Computing Machinery, 2008).
35. 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.* **31**, 141–146 (2022).
36. 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 (2023).
37. 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).
38. 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).
39. 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.* https://doi.org/10.3389/fimmu.2020.01951 (2020).
40. 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).
41. 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).
42. Riechmann, L., Clark, M., Waldmann, H. & Winter, G. Reshaping human antibodies for therapy. *Nature* **332**, 323–327 (1988).
43. Saengjaruk, P. et al. Diagnosis of human leptospirosis by monoclonal antibody-based antigen detection in urine. *J. Clin. Microbiol.* **40**, 480–489 (2002).
44. Sang, Z., Xiang, Y., Bahar, I. & Shi, Y. Llamanade: an open-source computational pipeline for robust nanobody humanization. *Structure* **30**, 418–429.e3 (2022).
45. Moutel, S. et al. NaLi-H1: a universal synthetic library of humanized nanobodies providing highly functional antibodies and intrabodies. *eLife* https://doi.org/10.7554/eLife.16228.001 (2016).
46. Xiang, Y. et al. Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2. *Science* **370**, 1479–1484 (2020).
47. Padlan, E. A. A possible procedure for reducing the immunogenicity of antibody variable domains while preserving their ligand-binding properties. *Mol. Immunol.* **28**, 489–498 (1991).
48. Roguska, M. A. et al. Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proc. Natl Acad. Sci. USA* **91**, 969–973 (1994).
49. Vanderhaegen, S. et al. Structure of an early native-like intermediate of β2-microglobulin amyloidogenesis. *Protein Sci.* **22**, 1349–1357 (2013).
50. Abanades, B. et al. ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. *Commun. Biol.* **6**, 575 (2023).
51. Domanska, K. et al. Atomic structure of a nanobody-trapped domain-swapped dimer of an amyloidogenic β2-microglobulin variant. *Proc. Natl Acad. Sci. USA* **108**, 1314–1319 (2011).
52. Schoof, M. et al. An ultrapotent synthetic nanobody neutralizes SARS-CoV-2 by stabilizing inactive Spike. *Science* **370**, 1473–1479 (2020).
53. Honegger, A. & Pluckthun, A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J. Mol. Biol.* **309**, 657–670 (2001).
54. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
55. Clavero-Álvarez, A. et al. Humanization of antibodies using a statistical inference approach. *Sci. Rep.* **8**, 14820 (2018).
56. Jiang, J. et al. Preclinical safety profile of disitamab vedotin: a novel anti-HER2 antibody conjugated with MMAE. *Toxicol. Lett.* **324**, 30–37 (2020).
57. Deeks, E. D. Disitamab vedotin: first approval. *Drugs* **81**, 1929–1935 (2021).
58. Aprile, F. A. et al. Rational design of a conformation-specific antibody for the quantification of Aβ oligomers. *Proc. Natl Acad. Sci. USA* **117**, 13509–13518 (2020).
59. 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).
60. 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).
61. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **40**, 1114–1122 (2022).
62. 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).
63. 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).
64. Tsuruta, H. et al. AVIDa-hIL6: a large-scale VHH dataset produced from an immunized alpaca for predicting antigen-antibody interactions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.03329 (2023).
65. Li, X. et al. Comparative analysis of immune repertoires between Bactrian camel's conventional and heavy-chain antibodies. *PLoS ONE* **11**, e0161801 (2016).
66. McCoy, L. E. et al. Molecular evolution of broadly neutralizing llama antibodies to the CD4-binding site of HIV-1. *PLoS Pathog.* **10**, e1004552 (2014).
67. Xiang, Y. et al. Integrative proteomics identifies thousands of distinct, multi-epitope, and high-affinity nanobodies. *Cell Syst.* **12**, 220–234.e9 (2021).
68. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).
69. Zeghidour, N., Luebs, A., Omran, A., Skoglund, J. & Tagliasacchi, M. SoundStream: an end-to-end neural audio codec. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2107.03312 (2021).
70. Yu, J. et al. Vector-quantized image modeling with improved VQGAN. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2110.04627 (2022).
71. Kaiser, Ł. et al. Fast decoding in sequence models using discrete latent variables. *Proc. Mach. Learn. Res.* **80**, 2390–2399 (2018).

72. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (eds Burstein, J. et al.) 4171–4186 (ACL, 2019).
73. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Annual Conference on Neural Information Processing Systems 2019* (eds Wallach, H et al.) 8024–8037 (Curran Associates, 2019).
74. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations* (ICLR, 2015).
75. Lefranc, M. P. & Lefranc, G. Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions. *Biomedicines* https://doi.org/10.3390/biomedicines8090319 (2020).
76. Schmitz, S., Soto, C., Crowe, J. E. Jr & Meiler, J. Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires. *mAbs* **12**, 1758291 (2020).
77. Gao, S. H., Huang, K., Tu, H. & Adler, A. S. Monoclonal antibody humanness score and its applications. *BMC Biotechnol.* **13**, 55 (2013).
78. Abhinandan, K. R. & Martin, A. C. R. Analyzing the 'degree of humanness' of antibody sequences. *J. Mol. Biol.* **369**, 852–862 (2007).
79. Raybould, M. I. J. et al. Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Res.* **48**, D383–D388 (2020).
80. Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351–371 (1973).
81. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilities of residues in proteins. *PLoS ONE* **8**, 80635 (2013).
82. Chen, H. & Zhou, H.-X. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.* **33**, 3193–3199 (2005).
83. Esposito, L., Vitagliano, L., Zagari, A. & Mazzarella, L. Pyramidalization of backbone carbonyl carbon atoms in proteins. *Protein Sci.* **9**, 2038–2042 (2000).
84. Ramon, A. et al. AbNatiV 1.0. *Zenodo* https://doi.org/10.5281/zenodo.10171047 (2023).

Acknowledgements

P.S. is a Royal Society University Research Fellow (grant no. URF\R1\201461). We acknowledge funding from UK Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (grant no. EP/X024733/1, an ERC starting grant to P.S. underwritten by UKRI) and from an Isaac Newton Trust/Wellcome Trust ISSF/University of Cambridge Joint Research grant (no. MBAG/624 RG89305 to P.S.). M. Ali is supported by a Harding Distinguished Postgraduate Scholarship and M.G. is a Yusuf Hamied Graduate Scholar.

Author contributions

P.S. conceived and supervised the project. A.R. developed the deep learning architecture with the guidance of M.G. and P.S.; A.R. parsed and collected the training data with the help of A.S. and K.D.; and A.R. built the humanization pipeline and designed the humanized variants. M. Ali, M. Atkinson and X.X. produced the nanobody variants and carried out wet-laboratory experiments. C.V. and S.R. produced β_2-microglobulin and provided expert advice. A.R. and P.S. wrote the first version of the paper. All authors analysed data and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

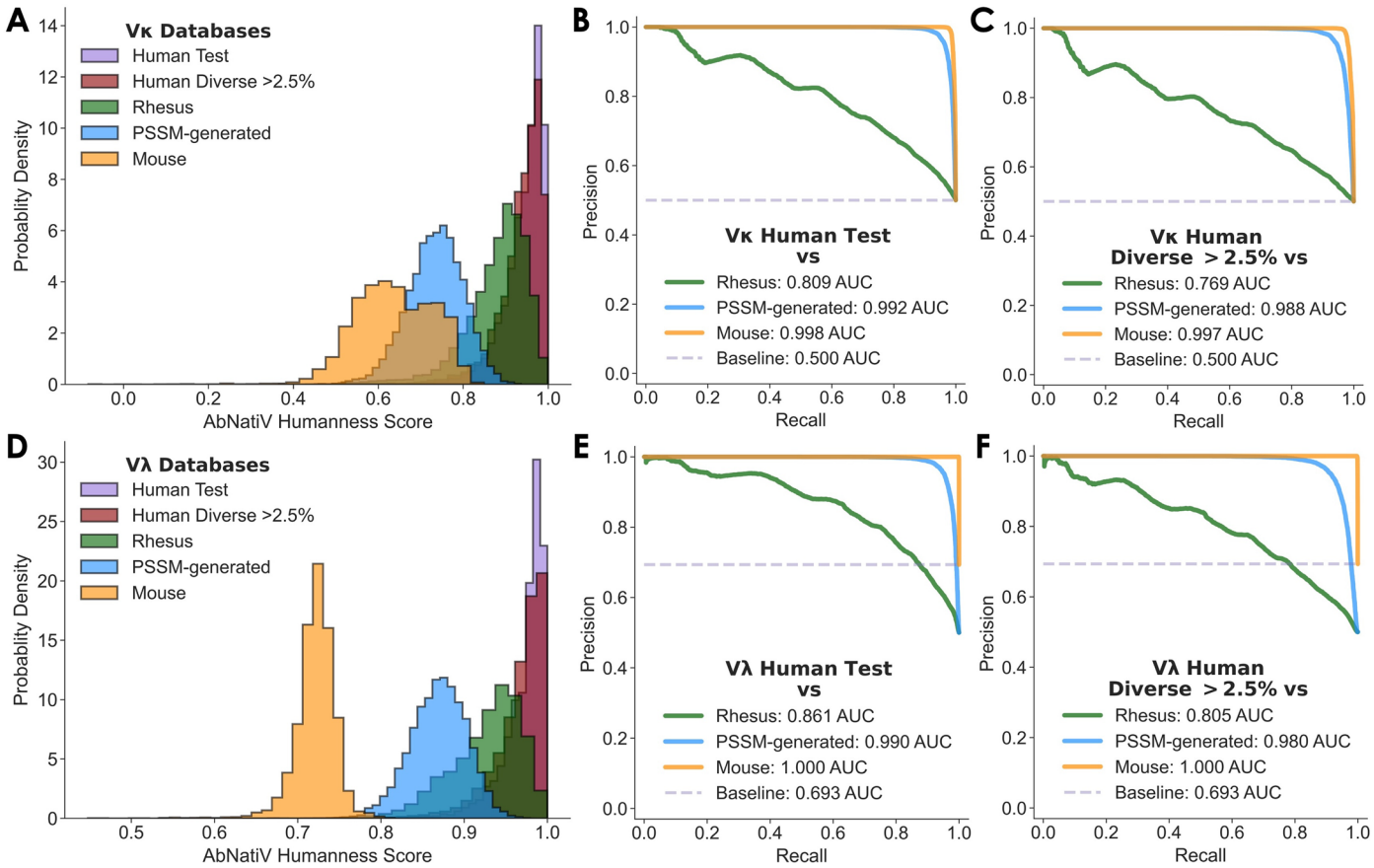
Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00778-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00778-3.

Correspondence and requests for materials should be addressed to Pietro Sormanni.

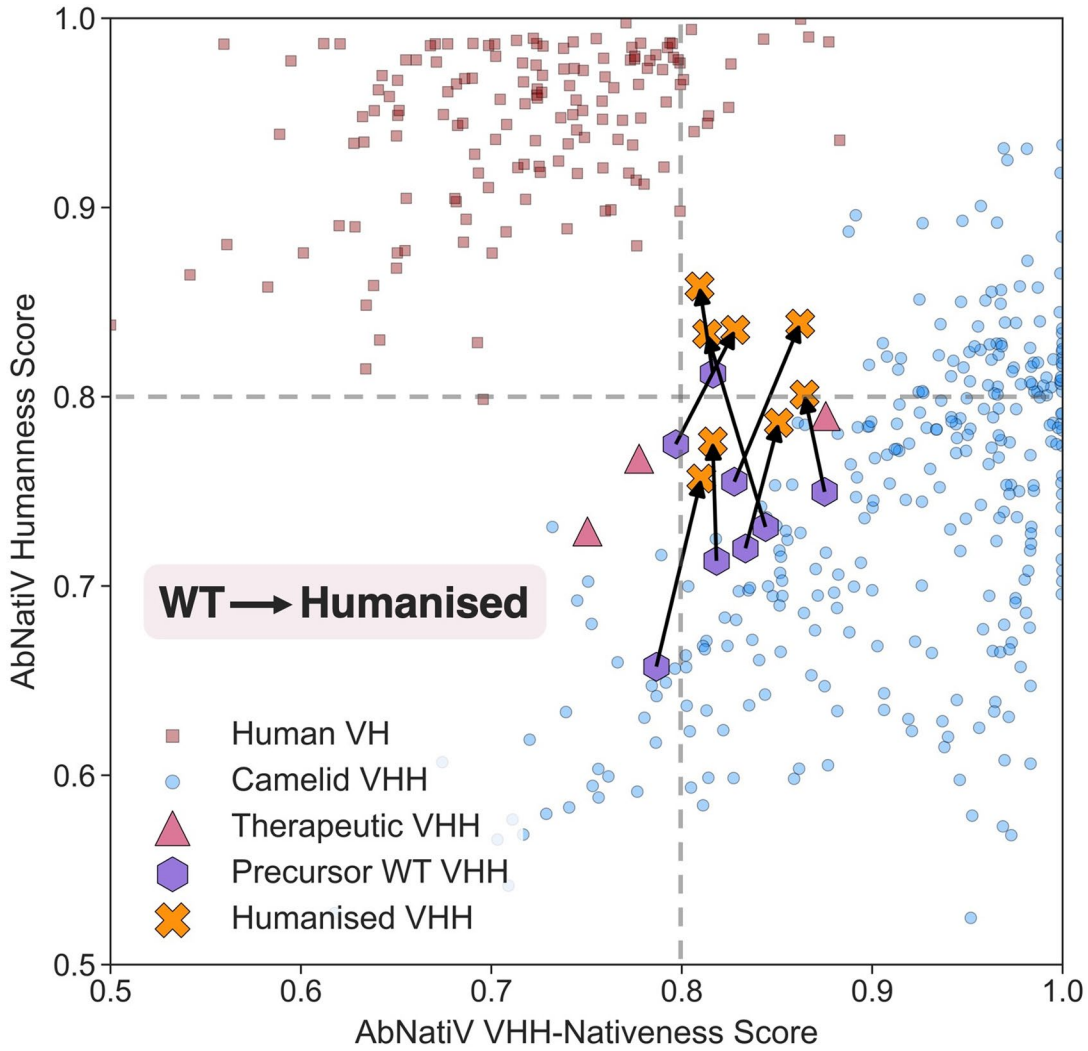
Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-023-00778-3



Extended Data Fig. 1 | Performance on V_κ and V_λ sequence classification. (**a**, **d**) The AbNatiV-humanness score distributions of the Human Test (purple), Human Diverse >2.5% (red), Rhesus (green), PSSM-generated (blue), and Mouse (orange) V_κ (**A**) and V_λ (**D**) 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. The Human Diverse >2.5% dataset is made of sequences from the Test and BioPhi datasets with a sequence identity difference of 2.5% from their respective closest sequence of the corresponding Training set (see Methods). Each dataset contains 10,000 sequences except Human Diverse >2.5% which contains 10,490 sequences for V_κ, and 10,459 for V_λ. (**b**, **c**) Plots of the PR curves computed to represent the ability of AbNatiV to distinguish the V_κ Human Test set (**B**) or Human Diverse >2.5% (**C**) from the other datasets (see legend, which also reports the area under the curve). (**e**, **f**) Same PR plots but for the V_λ model. The corresponding ROC curves are given in Supplementary Fig. 6c–f. The baseline (dashed line) corresponds to the performance that a random classifier would have with the Mouse dataset.

https://doi.org/10.1038/s42256-023-00778-3



Extended Data Fig. 2 | Combining AbNatiV humanness and V_HH-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 the three nanobody therapeutics Envafolimab, Caplacizumab, and Rimteravimab (in pink), and 8 WT nanobodies (in purple) with their humanized counterpart (in orange) (45). 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 to improve readability. To provide a reference background distribution, 300 randomly selected human VH sequences and 300 camelid VHH sequences are plotted. The cluster of human sequences that score relatively well in VHH-nativeness derive from the IGHV-3 germline gene (Supplementary Fig. 23), consistent with the genetic origin of natural camelid nanobodies (48).

https://doi.org/10.1038/s42256-023-00778-3

Extended Data Table 1 | Evaluation of the PR classification and reconstruction tasks for human V_κ light-chain sequences

V_κ	Classification (PR-AUC)						Reconstruction accuracy	
	Rhesus vs		Mouse vs		PSSM-generated vs		Human Test (T)	Human Diverse >2.5% (D)
	T	D	T	D	T	D		
AbNatiV	0.809	0.769	0.998	0.997	0.992	0.988	0.982	0.979
OASis (relaxed)	0.734	0.744	0.993	0.993	0.959	0.961	N/A	N/A
Sapiens	0.848	0.860	0.993	0.993	0.989	0.989	0.935	0.939

The assessment is carried out for AbNatiV trained on human V_κ sequences (first row) and other computational approaches that can assess humanness (other rows). The first six columns report the PR-AUC (curves shown in Extended Data Fig. 1b,c and Supplementary Fig. 9a–d), assessing the ability of the models to separate sequences in the Human Test (T) or the Human Diverse >2.5% (D) sets from those from mouse, rhesus, and PSSM-generated (see column headers). 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. Many sequences of the D datasets belong to the Sapiens training set. See ROC results in Supplementary Table 2.

https://doi.org/10.1038/s42256-023-00778-3

Extended Data Table 2 | Evaluation of the PR classification and reconstruction tasks for human V_λ light-chain sequences

V_λ	Classification (PR-AUC)						Reconstruction accuracy	
	Rhesus vs		Mouse vs		PSSM-generated vs		Human Test (T)	Human Diverse >2.5% (D)
	T	D	T	D	T	D		
AbNatiV	0.861	0.805	1.000	1.000	0.990	0.980	0.983	0.978
OASis (relaxed)	0.818	0.822	1.000	0.999	0.958	0.955	N/A	N/A
Sapiens	0.876	0.877	0.999	0.999	0.978	0.967	0.930	0.932

The assessment is carried out for AbNatiV trained on human V_λ sequences (first row) and other computational approaches that can assess humanness (other rows). The first six columns report the PR-AUC (curves shown in Extended Data Fig. 1d,e and Supplementary Fig. 9e–h), assessing the ability of the models to separate sequences in the Human Test (T) or the Human Diverse >2.5% (D) sets from those from mouse, rhesus, and PSSM-generated (see column headers). 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. Many sequences of the D datasets belong to the Sapiens training set. See ROC results in Supplementary Table 3.

https://doi.org/10.1038/s42256-023-00778-3

Extended Data Table 3 | Performance on the classification of antibody therapeutics

Method	Human vs Non-Human	
	ROC AUC	PR AUC
AbNatiV	0.979	0.971
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

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 (that is, trained either on VH, V_κ, or V_λ, 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.

https://doi.org/10.1038/s42256-023-00778-3

Extended Data Table 4 | Evaluation of the PR classification and reconstruction tasks for camelid V_HH sequences

V_HH	Classification (PR-AUC)								Reconstruction accuracy	
	Human Test vs		Mouse vs		Rhesus vs		PSSM-generated vs		Camelid Test (T)	Camelid Diverse >5% (D)
	T	D	T	D	T	D	T	D		
AbNatiV	0.983	0.961	0.995	0.987	0.992	0.980	0.942	0.893	0.954	0.954
AbLSTM retrained	0.956	0.900	0.988	0.969	0.983	0.956	0.916	0.839	0.847	0.846

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 PR curves (shown in Fig. 4c and Supplementary Fig. 12**)**, assessing the ability of the models to separate sequences in the Camelid Test (T) or Human Diverse >5% (D) sets from those from human, mouse, rhesus, and PSSM-generated (see column headers). The Camelid Diverse >5% dataset is used as a control to specifically assess the ability to generalize to sequences distant from those in the training set. The last two columns quantified the ability of each model to reconstruct camelid sequences in each dataset (column header). Corresponding ROC results are in Supplementary Table 4.

https://doi.org/10.1038/s42256-023-00778-3

Extended Data Table 5 | Nanobody sequences experimentally tested

Nanobody	Sequence	Theoretical MW	Observed MW
Nb24 WT	QVQLQESGGGSVQAGGSLRLSCAASGYTDSRYCMAWFRQAPG KEREWVARINSGRDITYYADSVKGRFTFSQDNAKNTVYLQMD SLEPEDTATYYCATDIPLRCRDIVAKGGDGFRYWGQGTQVTVS SLEHHHHHH*	15213.67	15209
Nb24 Enhanced sampling	EVQLLESGGGLVQPGGSLRLSCAASGYTDSRYCMAWFRQAPG KEREWVARINSGRDITYYADSVKGRFTFSRDNSKNTVYLQMDS LRPEDTAVYYCATDIPLRCRDIVAKGGDGFRYWGQGTQVTVSS LEHHHHHH*	15320.94	15317
Nb24 Enhanced sampling (+buried)	EVQLLESGGGLVQPGGSLRLSCAASGYTDSRYCMAWFRQAPG KEREWVARINSGRDITYYADSVKGRFTVSRDNSKNTVYLQMDS LRPEDTAVYYCATDIPLRCRDIVAKGGDGFRYWGQGTQVTVSS LEHHHHHH*	15272.90	15268
Nb24 Exhaustive sampling	EVQLVESGGGLVQPGGSLRLSCAASGYTDSRYCMAWFRQAPG KGLEWVARINSGRDITYYADSVKGRFTFSRDNAKNTVYLQMD SLRPEDTAVYYCATDIPLRCRDIVAKGGDGFRYWGQGTQVTVS SLEHHHHHH*	15175.83	15172
Nb24 Exhaustive sampling (+buried)	EVQLVESGGGLVQPGGSLRLSCAASGYTDSRYCMAWFRQAPG KGLEWVARINSGRDITYYADSVKGRFTISRDNAKNTVYLQMDS LKPEDTAVYYCATDIPLRCRDIVAKGGDGFRYWGQGTLVTVSS LEHHHHHH*	15098.82	15095
Nb24 Frequency & structure-based	QVQLVESGGGLVQPGGSLRLSCAASGYTDSRYCMAWFRQAPG KGLEWVARINSGRDITYYADSVKGRFTISRDNAKNTLYLQMNS LRAEDTAVYYCARDIPLRCRDIVAKGGDGFRYWGQGTLVTVSS LEHHHHHH*	15167.93	15164
mNb6 WT	MEVQLVESGGGLVQAGGSLRLSCAASGYIFGRNAMGWYRQAP GKERELVAGITRRGSITYYADSVKGRFTISRDNAKNTVYLQMN SLKPEDTAVYYCAADPASPAYGDYWGQGTQVTVSSHHHHHH*	13755.25	13752
mNb6 Enhanced sampling	EVQLVESGGGLVQPGGSLRLSCAASGYIFGRNAMGWYRQAPG KERELVAGITRRGSITYYADSVKGRFTISRDNAKNTVFLQMNSL RPEDTAVYYCAADPASPAYGDYWGQGTQVTVSSHHHHHH*	13662.10	13660
mNb6 Enhanced Sampling (+buried)	EVQLVESGGGLVQPGGSLRLSCAASGYIFGRNAMGWVRQAPG KGREWVSGITRRGSITYYADSVKGRFTISRDNAKNTVYLQMNS LRPEDTAVYYCAADPASPAYGDYWGQGTQVTVSSHHHHHH*	13631.06	13628
mNb6 Exhaustive sampling	EVQLVESGGGLVQPGGSLRLSCAASGYIFGRAMGWYRQAPGK GLEWVAGITRRGSITYYADSVKGRFTISRDNAKNTVFLQMDSL RPEDTAVYYCAADPASPAYGDYWGQGTLVTVSSHHHHHH*	13606.07	13604
mNb6 Exhaustive sampling (+buried)	EVQLVESGGGLVQPGGSLRLSCAASGYIFGRNAMGWVRQAPG KGLEWVAGITRRGSITYYADSVKGRFTISRDNAKNTVYLQMDS LRPEDTAVYYCAADPASPAYGDYWGQGTLVTVSSHHHHHH*	13558.03	13556
mNb6 Frequency & structure-based	QVQLVESGGGLVQPGGSLRLSCAASGYIFGRNAMGWVRQAPG KGLEWVAGITRRGSITYYADSVKGRFTISRDNAKNTLYLQMNS LRAEDTAVYYCARDPASPAYGDYWGQGTLVTVSSHHHHHH*	13629.16	13627

Sequences of the WT nanobodies Nb24 and mNb6 and their humanised variants as used in the wet-lab experiments. A PelB signal sequence was present at the N-terminus of all nanobodies, but this is cleaved upon secretion and hence it is not part of the final protein. All humanised designs are done with AbNatiV except for the Frequency & structure-based designs, which are done with the Llamanade software (45). The theoretical MW is calculated from the amino acid sequence assuming reduced di-sulphide bonds, observed MW is measured with LC-MS. Nb24 variants have two disulfide bonds and mNb6 have one. Therefore, a difference of −4 Da and −2 Da respectively for Nb24 and mNb6 variants is expected between theoretical and observed MWs.

nature portfolio

Corresponding author(s): Pietro Sormanni

Last updated by author(s): Nov 23, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected from the sources detailed in Supplementary Table 1 without the need of any software.
Data analysis	Antibody Fv sequences were aligned as described in the method section, using the open-source ANARCI (v. 1.b) software. PyTorch (1.13.1) and PyTorchLightning (0.7.3) were used to monitor the training/validation performances. ImmuneBuilder (1.0.1) was used to generate antibody structures. Llamanade (webserver: http://35.208.211.136/ on 4th of July 2023) was used to humanise Nb24 and mNb6 nanobodies. The Expasy ProtParam toll (webserver: web.expasy.org/protparam/) was used to compute the extinction coefficients of the expressed antibodies. The UNIFI software was used to analyse the LCMS spectra. All the code we have developed and used in the manuscript is made available open source at https://gitlab.developers.cam.ac.uk/ch/sormanni/abnativ (https://doi.org/10.5281/zenodo.10171047).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the large training/validation/testing datasets needed to train and evaluate AbNatiV are available online in the AbNatiV GitLab at https://gitlab.developers.cam.ac.uk/ch/sormanni/abnativ/-/tree/main/datasets?ref_type=heads (release 1: https://doi.org/10.5281/zenodo.10171047). Details and sources of these datasets are presented in Supplementary Table 5. Smaller datasets required to analyse the therapeutic classification, ADA correlation, and VHH grafting studies are compiled in the Supplementary Datasets file, with details and sources included in the legends. All the sequences tested in vitro are provided in Extended Data Table 5.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The method is trained, validated and tested on the number of sequences reported in the material and method and detailed in Supplementary Table 5. For the VHH model, we trained it with all the VHH camelid repertoires available in the literature at the time. For the human models, we provide a sample size study in Supplementary Fig. 18A (see the Discussion section).
Data exclusions	Sequences that were incomplete, or that did not include ultra-conserved Cys residues were discarded as described in the Methods section.
Replication	All the binding experiments were replicated 3 times. The thermal stability experiments for Nb24 were replicated 3 times. The thermal stability experiments for mNb6 were replicated twice.
Randomization	Splitting between the train/validation/test datasets was fully randomised.
Blinding	All the testing datasets were never seen by the model. The hyper-parameters of the model were selected with the validation dataset.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All the antibodies used in the humanisation study were expressed by us. Details and sequences for each antibody are presented Extended Data Fig. 5.
Validation	All the details of production, purification and characterisation are presented in the Method section.