An Antibody Developability Triaging Pipeline Exploiting Protein Language Models

James Sweet-Jones[1] and Andrew C.R. Martin[1,2]

[1]Institute of Structural and Molecular Biology,
Division of Biosciences, University College London,
Darwin Building, Gower Street,
London, WC1E 6BT, UK

November 29, 2024

[2]**Corresponding author**

eMail: andrew@bioinf.org.uk / andrew.martin@ucl.ac.uk

Telephone: +44 207 679 7034

Fax: +44 207 679 7193

Abstract

Therapeutic monoclonal antibodies are a successful class of biologic drugs that are frequently selected from phage display libraries and transgenic mice that produce fully human antibodies. However, binding affinity to the correct epitope is necessary, but not sufficient, for a mAb to have therapeutic potential. Sequence and structural features affect the developability of an antibody, which influences its ability to be produced at scale and enter trials, or can cause late-stage failures. Using data on paired human antibody sequences, we introduce a pipeline using a machine learning approach that exploits protein language models to identify antibodies which cluster with antibodies that have entered the clinic and are therefore expected to have developability features similar to clinically acceptable antibodies, and triage out those without these features. We propose this pipeline as a useful tool in candidate selection from large libraries, reducing the cost of exploration of the antibody space, and pursuing new therapeutics.

1 Introduction

Monoclonal antibodies (mAbs) have been shown to be a successful class of biologic drugs which have potential to treat a wide variety of diseases owing to their ability to target a specific antigen, and therefore potentially any step in a disease pathway [1, 2]. At the time of writing, at least 130 mAbs have received regulatory approval from the U.S. Food and Drug Administration or the European Medicines Agency with at least 42 being considered as 'fully-human', either from transgenic mice, phage display libraries, or cloned from recovering patients [3, 4, 5]. The annual growth of this sector has increased at between 20% and 30% per year [6, 7], and is likely to continue to grow as interest increases in the use of antibodies to target previously undruggable targets [8]. Despite this, throughout the clinical pipeline for the development of new mAbs, there is a high risk of failure, causing costly discontinuation from trials [9].

Simultaneously, efforts in single cell sequencing techniques have been applied to understand how the antibody repertoire functions and changes over time at the level of single B cells [10, 11, 12, 13]. This has given researchers the ability to generate dense digital libraries of paired variable heavy (V_H) and variable light (V_L) human antibody sequences that vastly outnumber previous databases resulting from sequence or structural data (KabatMan [14], IMGT [15], SAbDAb [16] AbDb [17] and EMBLIg (abybank.org/emblig/)). Online repositories including the Observed Antibody Space (OAS) [18], cAb-Rep [19] and BRepertoire [20] allow researchers access to these resources.

With the generation of these *in silico* databases, efforts to develop screening statistics to identify sequences with physical characteristics similar to approved therapeutics has become a driver in the field. Usually, these have been based on antibody developability, which is loosely defined as an antibody's intrinsic ability to be produced on an industrial scale, to maintain reasonable stability in long-term storage and in patients, and to be safely tolerated by the patient [21, 22]. Such considerations have now become important in the early stages of drug screening to select the best quality candidates and avoid costly late-stage failures [1]. Furthermore, developability is important, but does not guarantee success at clinical trials, where candidates may face discontinuation for safety or efficacy reasons. Identifying factors important in determining success at clinical trials has also eluded researchers.

Physicochemical features, including surface charged patches, surface hydrophobic patches, low thermostability, and post-translational modification sites that introduce heterogeneity, have become associated with poor antibody developability [23]. Those features that compromise the stability of the antibody can cause unfolding, increase the propensity to aggregate in solution and can increase immunogenicity [24, 25]. At the lead candidate stage, well-defined experimental assays for measurement are important in the selection of a final lead [26, 27]. However, it has become useful to predict these features at an earlier stage using computational means. To this end,

sequence-based statistics have been developed based on these features and are available for use in drug discovery pipelines including the Developability Index [28, 29], AbPred [30], and, more recently, the Therapeutic Antibody Profiler (TAP) [31] and Therapeutic Antibody Developability Analysis (TA-DA Score) [32]. However, these tools can fall short in identifying leads from large libraries of data, requiring computationally expensive 3D modelling, or only taking one antibody at a time, which is usually expected already to be a potential lead candidate.

In order to take advantage of the wealth of data now available, the field has also turned to machine learning as a new avenue of exploration [1, 33, 34]. For protein sequences to be suitable inputs for machine learning problems, it is necessary to encode them numerically. Previously, this has been done by using evolutionary or physicochemical and structural features [35, 36, 37], and simple regression models to identify features of high importance, or to predict features from the sequence as done in AbPred [30]. Negron *et al*. [32] expanded on this work and identified previously mentioned characteristics, including hydrophobicity (assessed by hydrophobic interaction chromatography), thermostability (Tm, assessed by differential scanning fluorimetry) and aggregation (assessed by cross-interaction chromatography) that were associated with the identification of clinically acceptable mAbs. Furthermore, this work has demonstrated an ability to separate clinical antibody sequences from antibody repertoires and to assign a developability score based on these features as part of their 'TA-DA' score.

A newer method of encoding protein sequences is using 'protein language models' [38]. These are deep learning encoders trained on the relationships between residues in a sequence using millions of sequences. The results give dense numerical representations of sequences that may then be used as training data for machine learning models [39]. In this study, we hypothesized that, rather than directly predicting physical properties related to developability, antibodies with developable traits may be selected by encoding them using protein language models and comparing the encoded antibodies with encoded sequences of current clinical mAbs (i.e. approved, discontinued and in-development mAbs). Our goal is then to build a high-throughput triaging pipeline exploiting preliminary simple physicochemical screening followed by machine learning using protein language models which may be used to select antibodies with good developability characteristics from large libraries.

2 Results

2.1 Simple physicochemical properties of clinical and library antibodies

As a first step, we looked at using physicochemical properties to attempt to identify antibodies with clinically acceptable properties in a set of library antibodies. The aim was to see whether the clinical mAbs have a restricted distribution of these properties compared with antibodies from a

library, similar to the approach used by Raybould *et al*. [31], except here, we currently use only sequence statistics that can be calculated quickly without high computational expense.

A dataset was collated consisting of paired V_H and V_L sequences of clinical stage human mAbs (*n*=144) from the October 2021 release of TheraSabDab [3] marked as 'Whole mAb' (Supplementary Table 1) and 10,000 paired sequences randomly selected from the OAS online repertoire repository (accessed January 2022) [18] (Supplementary Table 2). We refer to this set of sequences from OAS as our 'library'.

Physicochemical properties, including predicted *ΔG* of unfolding [40], iso-electric point (pI) [41] and CDR-H3 loop length [42] were calculated. Using an unpaired non-parametric t-test, it was observed that there were statistical differences in the CDR-H3 length and in the predicted *ΔG* of unfolding for concatenated V_H and V_L chains between therapeutic and library antibodies (Table 1 and Supplementary Table 3). While this demonstrates a difference between human repertoire antibodies and what is found in the clinical mAb dataset, the mean values are relatively similar in the two datasets making it difficult to use this as an approach to identify antibodies with clinically acceptable developability characteristics, although it can be used to reject clear outliers.

2.2 Identifying Clinical-like Antibodies from Repertoires Using Unsupervised Learning

An unsupervised learning model was proposed as an approach to identify library antibodies with clinical-antibody-like properties. It would be expected that clinical mAbs would cluster in some *N*-dimensional space and that repertoire antibodies with similar properties would be positioned close to the clinical mAbs. To train an unsupervised learning method, the library and clinical V_H and V_L sequences were padded according to the Chothia numbering scheme, then independently encoded with various language models: ESM [39], AbLang [43], Sapiens [44] and AntiBERTy [45]. The encodings generated 130,048 features per paired V_H/V_L sequence. All language models had a similar performance for this task, with AntiBERTy somewhat out-performing the other methods (data not shown).

Various unsupervised machine learning models were tested: linear Principal Component Analysis (PCA) [46], kernel PCA [46], 2D 't-distributed Stochastic Neighbour Embedding' (t-SNE) [47] and 'Uniform Manifold Approximation and Projection' (UMAP) [48] (Figure 1a). These algorithms demonstrate how library antibodies are positioned against clinical mAbs also encoded with the AntiBERTy language model. For the linear PCA, t-SNE or UMAP, data were arranged into discrete groups of antibodies which are dictated by V_H and V_L gene germline pairing (Figure 1b). However, Non-linear PCA with a radial basis kernel function (γ = 500, Supplementary Figure S1), when viewing the first two principal components, gave a useful pattern of clustering where library antibodies form a radial pattern with clinical mAbs positioned around the origin (Figure 1a). This

was also true of a held back dataset of human-derived clinical mAbs (*n*=203) named with the 2016 and 2022 naming conventions [49] in which the source infix was removed, and therefore human mAbs could not be identified using the '-umab but not -zumab' approach used to identify human mAbs with the earlier naming schemes (Supplementary Table 4). These held-back antibodies were positioned close to the original dataset of human-clinical mAbs (Figure 2). This led us to conclude that repertoire antibodies which are positioned close to clinical mAbs may be likely to share the developability properties necessary and should be taken forward for potential development.

Cutoffs were then established to select the repertoire antibodies which cluster with the clinical mAbs in order to extract them. An ellipse function was used in which the principal component with the greater range for clinical mAbs was taken to be the major axis, and the lesser range as the minor axis. Z-score thresholds (the number of standard deviations away from the mean) along the two principal components of the clinical mAbs were used to select where the extremes of the ellipse should be placed. The Z-score thresholds were optimized by measuring the proportion of the clinical mAbs captured by the ellipse against the proportion of the library antibodies also captured in the same ellipse. It was expected that, since the spread of antibodies was even across the first two principal components of the PCA, roughly equal proportions of both groups would be captured. This was done using all human clinical mAbs (Figure 3a) and with only approved human mAbs (Figure 3b).

Comparing Figures 3a and 3b, it can be seen that the bars for the OAS (library) antibodies are consistently lower when the Z-scores are based on the approved antibodies than they are when based on the clinical (i.e. approved, discontinued and in-development) antibodies. This indicates that the approved antibodies occupy a tighter distribution than the clinical antibodies. While it is obvious that the approved antibodies will be a subset of the clinical antibodies, it is less obvious that they will form a tighter cluster in this projection of the AntiBERTy-encoded parameter space. This led us to conclude that there may be characteristics of the approved antibodies identified by the protein language model that would allow them to be separated from the antibodies that were discontinued.

2.3 Using Supervised Machine Learning to Distinguish Approved and Discontinued Clinical antibodies

It is evident that having suitable developability profiles alone is not sufficient for an antibody to succeed in clinical trials and the clinical dataset used to identify library antibodies with properties similar to clinical mAbs contained discontinued antibodies. There are many reasons why an antibody could fail in clinical trials. Some of these are intrinsic to the sequence (immunogenicity, developability), while others are target-specific (binding affinity, nature of the epitope, on- or off-target side-effects, etc.) [1, 9, 50]. However, given the differences observed above, we considered it worthwhile to attempt to train a predictor that might be able to identify drug-like antibodies that

are likely to succeed in the clinic. We assembled a dataset of the V_H and V_L amino acid sequences for 115 approved and 150 discontinued antibodies from the TheraSabDab [3] (Supplementary Table 5).

Unlike the comparison of human clinical mAbs and library antibodies, which had statistical differences in their physicochemical properties (Table 1), there were no statistical differences in any of the basic physicochemical properties between approved and discontinued antibodies using an unpaired t-test. This included the G score [51] used as a method for predicting immunogenicity (Table 2). The largest quantitative difference was that the discontinued antibodies had a lower mean length for the CDR-H3 loop (Supplementary Table 6). The lack of statistical differences is perhaps not surprising given that both approved and discontinued antibodies will almost certainly have undergone a developability assessment and possibly optimisation before entering clinical trials. It was also seen that the approved and discontinued groups had similar proportions of V_H and V_L V-gene germline pairings (Supplementary Figure S2).

As before, the V_H and V_L sequences for each antibody were padded and aligned using the Chothia numbering scheme and the sequences were then encoded with a selection of general protein (ESM [39]) and antibody-specific (AntiBERTy [45], AbLang [43], Sapiens [44]) protein language models. The encodings of the paired V_H and V_L sequences were concatenated and treated as a single set of data points per antibody. The encoded antibody sequences were used to train a selection of 15 supervised machine learning classifiers (see the Methods and Supplementary File: Supplementary ML.pdf). Models were trained with ten-fold cross validation (CV) and model performance was evaluated using the mean Matthews' Correlation Coefficient (MCC) [52] of the predictions of the test split of each fold. F-regression was used as a method of feature selection, by selecting the *k* features most correlated with the labels where *k* was set to [1,10,50,100,500,1000,2500,5000,10000].

Generally performance across all classifiers was good (Supplementary Tables 7–10), but it was seen that the overall best performance was obtained for the AntiBERTy encodings, particularly when using F-regression for feature selection with *k* set to 2500. The top performing models were the Linear Support Vector Machine classifier (LinearSVC; MCC=0.8±0.08), Ridge Classifier (MCC=0.78±0.12) and Logistic regression (MCC=0.80±0.1) (See Figure 4). All were evaluated using a standard classification threshold of 0.5. The LinearSVC model was selected as the best model with a mean sensitivity of 0.86±0.10 and specificity of 0.93±0.05 across the 10 CV splits. In an attempt to improve the specificity further, this model was also assessed using a higher prediction threshold of 0.8. As expected, this resulted in a loss in sensitivity and an increase in specificity (Sn=0.57±0.17, Sp=0.99±0.03). This was accompanied by a decreased, but still respectable, MCC (0.64±0.11). See Table 3 and Supplementary Figure S3 which shows confusion matrices for the raw outputs of this model at both probabilities.

The location of the majority of selected features in the V_H or V_L sequences was then identified. Intuitively, it was expected that CDR-H3 would contain a high proportion of features, but this was not the case. Instead, a region in Framework 3 of the V_L chain had a high concentration of selected features, indicating that this region is highly important in how all the models have learned to distinguish these groups (Figure 5).

However, when a 'held back' dataset of therapeutics which have been approved (*n*=10) and discontinued (*n*=11) since the original access of TheraSabDab (Supplementary Table 11), the predictive score was found to be MCC=0.14 using a prediction threshold of 0.5, but this increased to MCC=0.51 with the higher prediction threshold was of 0.8 (Supplementary Figure S4a). A summary of results for the cross-validation and independent test sets is shown in Table 3. On the independent test set, the default prediction threshold results in a reasonable sensitivity and specificity as a result of numerous false positives; increasing the threshold to 0.8 improves the specificity accompanied by a small decrease in sensitivity resulting in a much improved MCC. Supplementary Figure S4b provides confusion matrices, MCC, sensitivity and specificity at prediction thresholds between 0.5 and 0.9.

It was surprising that such good performance was achieved using basic predictive models. Because it was previously shown that approved and discontinued antibodies did not show a statistically significant difference in features including isoelectric point, thermostability and CDR-H3 length (Table 2), it could be that other properties, such as immunogenicity [25], V-region germline family pairing or ability to access their targets are responsible for this ability to discriminate between these groups [45].

2.4 Assembling the pipeline to optimize performance and offer additional triaging.

The kernel PCA model (γ = 500) shown to separate library antibodies which are positioned close to clinical mAbs, and the LinearSVC model using 2500 features shown to separate approved and discontinued antibodies using the AntiBERTy language model encodings, were used to build a pipeline capable of selecting developable antibodies from an input library. The encoding only needs to take place once and can be carried forward between the two layers of models: unsupervised and supervised, respectively.

The complete pipeline attempts first to remove antibodies with obvious developability issues through physicochemical properties using Z-scores taken from the values of the approved antibody dataset ('Physicochemical Filtering' as given in Table 1, default Z=2); this saves computational time in numbering and encoding antibodies with the language model, as well as producing a better quality output. Antibodies with features typical of clinical mAbs are then selected from the unsupervised clustering of encoded antibodies using the ellipse function ('Layer 1') and are then

classified according to whether they are likely to pass clinical trials (‘Layer 2’). A user may enter a library of human antibodies and obtain entries from those that are most likely to be successful. A schematic of the pipeline is shown in Figure 6 where triaging stringency may be altered at each triaging step.

2.5 Testing on an example dataset demonstrates points of parameter tuning for optimized output

To illustrate the application of our pipeline, a library of 10,382 paired B-cell receptor (BCR) sequences taken from six healthy blood donors [54] was used as an example test dataset.

After physicochemical triaging, ‘Layer 1’ filtering is performed by performing PCA on the test data together with our ‘library’ antibodies from OAS and the previously used clinical dataset. The Z-score cutoffs are then calculated from the clinical dataset and the ellipse is generated and used to select antibodies from the test dataset.

Using decreasing Z-scores for the physicochemical triaging of the sequences reduced the number of antibodies entering ‘Layer 1’ (Table 4 and Supplementary Table 11). Similarly, decreasing the Z-score of the ellipse function in ‘Layer 1’ generally reduces the number of sequences taken forward to ‘Layer 2’ (Figure 7). However, since the clustering is performed and the ellipse is recalculated for each dataset there is some variation and, in one case (Table 4, no physicochemical filtering, ‘Layer 1’, Z-score= 1.0), there is a small rise in the number of antibodies compared with Z-score= 2.0. Increasing the prediction threshold used in ‘Layer 2’ also reduces the final number of selected antibodies.

As a comparison for the quality of antibodies output by the model, we checked the TAP score [31] for each antibody from the test BCR library. The TAP score is a developability score where an antibody with values for selected physicochemical properties that are seen within the clinical mAb dataset are given a perfect score of 0, and antibodies with increasing numbers of ‘red flags’ where the values are at the extremes of, or outside, the observed ranges are given negative scores. This is an indicator of developability, not whether an antibody is likely to be approved. It should be noted that this is a very different approach from our unsupervised machine learning: TAP relies solely on calculated or predicted physicochemical properties, while we use a subset of these properties only for a preliminary screen before using a clustering in high dimensional space obtained from a protein language model.

The median TAP score for antibodies in the Test BCR library was 0, which means that more than half of these antibodies were predicted to have no developability red flags. However, the minimum TAP score observed from the library was -110 indicating there are antibodies in the library with many developability red flags. From the data in Table 4, it is clear that setting the physicochemical

property filtering (PCF) in our approach to a more stringent Z-score (e.g. Z=0.5) had the major effect in removing antibodies with the most negative TAP scores from the output. Indeed with no PCF, neither the ‘Layer 1’ nor ‘Layer 2’ filtering removed the antibodies with TAP = -110. Similarly, as the ‘Layer 1’ stringency was increased, there was very little effect on the minimum, or the mean, TAP score. This is, perhaps, not surprising since the physicochemical properties on which this preliminary filtering is performed are somewhat similar to those exploited by the TAP score. However, the number of negative TAP scores does decrease as the ‘Layer 1’ filtering becomes more stringent (Table 4).

Again, because of the recalculation of the Z-scores and ellipse, there is one case in which the mean TAP score does not steadily progress closer to zero as the ‘Layer 1’ stringency is increased (Table 4, physicochemical filtering, Z-score=0.5).

It is also interesting to observe that, comparing the output of ‘Layer 1’ and ‘Layer 2’, the minimum and mean TAP scores improve. Given that ‘Layer 2 is predicting clinical success rather than developability, there is no reason to expect that this would be the case. Indeed, the percentage of antibodies with negative TAP scores retained after ‘Layer 2’ is larger than that after ‘Layer 1’ indicating that ‘Layer 2’ filtering is indeed detecting something different from developability.

3 Discussion

We have demonstrated the ability to triage library antibodies to find those with properties similar to currently available therapeutic mAbs. This has been achieved through a combination of preliminary filtering using physicochemical properties (to remove clearly outlying mAbs), with unsupervised and supervised machine learning. This demonstrates a useful tool in monoclonal antibody therapeutic discovery that may be applied to new and pre-existing paired human antibody libraries to identify potential clinical candidates with potential to pass clinical trials in order to avoid expensive late stage failures. Parameters of the pipeline at each step may be adjusted such that increased or reduced stringency filtering can produce a smaller (but more likely to be successful) or larger selection of antibodies. This pipeline can be used to identify antibodies with properties of therapeutic mAbs from large libraries [55, 56], to screen antibodies from transgenic animals following immunizations [57], or from human patients recovering from a condition of interest [5]. Using the pipeline in these contexts reduces the experimental work in finding an antibody which has properties suitable for use in the clinic [27].

This basic schematic for our pipeline allows for further optional triaging to be added at any point to give additional layers of stringency. The advantage of using these steps is a vastly reduced computation time. On an AU5000 GPU used in this study, the AntiBERTy encoding takes 0.06 seconds per V_H and V_L pair, making it suitable for the high-throughput analysis of libraries (compared with the 30 seconds required per antibody for the TAP Score web server [31]). While

the protein language models may be doing so implicitly, using additional features also opens up the possibility of using other explicit features including screening for immunogenicity [50] and known sequence liabilities such as post-translation modification sites [55, 58] and hydrophobic patches [24].

Direct comparisons of the performance of our method with the TAP score are not really possible. TAP relies on the distribution of a number of calculated and predicted physicochemical properties, some of which rely on a (predicted) structure of the antibody. Predicted properties can be compared with available experimental data. We only use physicochemical properties, calculated solely from sequence, as a preliminary screen to remove obvious outliers. The machine learning stages are based on a protein language model encoding that projects information (including implied structural features) into a high-dimensional space which is then reduced to a 2-dimensional space in which clinical antibodies are seen to cluster. Consequently, we do not directly predict properties related to developability and comparisons with published experimental data are not possible.

It is also worth noting that our approach is not simply suggesting that if the sequences are more similar to those of clinical-stage antibodies, they should have better developability. If that were the case, we could simply use a BLAST search. Rather, we exploit the encoding from the antibody-specific protein language model, AntiBERTy and it is well established that protein language model encoding of sequences relates a sequence to structural and lineage information on which it has been trained and thus captures other key information. These encodings are highly sensitive and can even predict the effect of single amino acid changes [71]. We have identified a 2D projection of the AntiBERTy encoding that clusters the clinical-stage antibodies and consequently we are looking at similarity of the protein language model encodings in those two principal components rather than sequence similarity *per se*.

Because human clinical antibodies clustered so closely in the kernel PCA, they must have similar features which have been encoded and recognized by the language model. The fact the clinical antibodies cluster near the origin suggests that they are developable largely because they are ‘ordinary’ antibodies which innately satisfy the required conditions. It is entirely possible that antibodies with very different properties could have therapeutic potential, but ultimately these would be higher-risk and consequently it is generally better to allow false negatives than false positives.

It is interesting to note that, using the TAP score, more than 25% of fully-human clinical-stage antibodies exhibit negative TAP scores. Our approach clusters all the clinical antibodies and the ellipse function (with default parameters) will capture all of these, including those that have TAP ‘red flags’. In other words, if input antibodies are found to be located within the same region of the

projection of the high-dimensional encoding, they are likely to have *sufficiently good* developability. We calculated the TAP scores for the 133 human clinical-stage antibodies and found that, while ~74% have a TAP score of zero (indicating no developability issues), the remainder have negative TAP scores as low as -30 (See Supplementary Figure S5).

While the unsupervised model ('Layer 1') groups together both approved and discontinued clinical antibodies, when these two groups are studied in a supervised context ('Layer 2'), it is possible to recognize differences between them. What can be concluded from this is that there are still features that separate them which are important for successfully completing clinical trials, and that the light chain Framework 3 seems to have a large contribution to these features. A drug may be discontinued from trials for efficacy reasons relating to bioavailability or binding to the target, safety reasons relating to the antigen or antibody (including immunogenicity) as well as marketing reasons [1, 9]. Since we showed that there are no statistical differences between the approved and discontinued groups for thermostability, pI or CDR-H3 properties, it is possible that the model is selecting features related to immunogenicity, or V_H/V_L germline gene pairing which may be related to stability [59]. The latter option could then be related to biases seen in the approved and discontinued datasets perpetuated by the lead candidate selection processes, but then, it has also been seen that the approved and discontinued dataset have similar proportions of V_H/V_L germline gene pairing, indicating other factors are being recognized in this region which are related to clinical trial success.

4 Conclusion

In conclusion, this work has demonstrated the ability to triage a library of antibodies to identify those with developability features similar to approved mAb therapeutics using language model encoding and applying them to both unsupervised and supervised machine learning. Furthermore, we demonstrate the ability to fine-tune the output in terms of quality by adjusting the thresholds of the models used to obtain the output. These tools aim to make use of previously curated and future antibody datasets to triage large datasets enabling faster and cheaper identification of potential lead candidates.

5 Methods

5.1 Data Collection

5.1.1 Human clinical-stage mAbs

Paired V_H and V_L sequences of therapeutic monoclonal antibodies (*n*=801) were downloaded from the October 2021 release of TheraSabDab [3]. Therapeutics marked as 'Whole mAb' were selected and identified as being fully human using the '-umab' suffix excluding instances of '-zumab' (humanized). Each therapeutic was checked for its source using the literature. This

resulted in a dataset of 143 antibodies: approved mAbs (*n*=31); discontinued mAbs (*n*=77) and in trials (*n*=35) (Supplementary Table 1). A further independent test dataset of human-derived clinical mAbs was acquired (*n*=203) using the 2016 naming convention in which the source infix was removed from the name and the 2022 naming convention using '-tug' for unmodified whole immunoglobulins and '-bart' for whole immunoglobulins with engineered amino acid changes in the constant domains [49] (Supplementary Table 4).

5.1.2 Library antibodies from OAS

The Observed Antibody Space database [18] was accessed in January 2022 and 34 libraries were downloaded totalling 88,274 paired sequences. 10,000 antibodies were selected randomly in order to create a training set for unsupervised learning (Supplementary Table 2).

5.1.3 Approved and discontinued mAbs

Clinical mAbs were obtained from the October 2021 release of the TheraSabDab database [3]. The V_H and V_L sequences of 115 approved antibody drugs and 156 discontinued drugs were collected. Seven drugs were excluded from the discontinued dataset as they were found to be discontinued for reasons not related to efficacy or safety. Edrecolomab was also moved from the approved dataset and the discontinued dataset, because it was later withdrawn for efficacy reasons [60]. The result of this is a dataset of 115 approved and 150 discontinued antibodies (Supplementary Table 5). Excluded sequences and reasons for their exclusion are found in Supplementary Table 12. A held back dataset of 21 therapeutics was taken from TheraSabDab accessed in October 2023 and not included in the original dataset (Supplementary Table 11).

5.1.4 Test BCR Dataset

The Test B cell receptor (BCR) sequence dataset [61] used to demonstrate the pipeline was downloaded from dx.doi.org/10.5281/zenodo.5146019. This dataset was obtained from six healthy blood donors who had their B cells isolated and FACS sorted by developmental stage. Transcripts from each individual cell were bar-coded making V_H/V_L pairing possible. Antibody V_H and V_L pairs were taken from B cells which shared the same bar-code where both an IGH and IGλ or IGκ chain was present. In cases where both IGλ and IGκ chains were present, the chain with the highest count number was taken as the V_L chain pair. No filtering based on the type or BCR developmental stage was performed. Individual amino acid sequences for frameworks and CDR loops were concatenated to give the full antibody variable domain sequence. In total, 10,382 paired antibodies were extracted.

5.2 Encoding H and L sequences with antibody language models

V_H and V_L sequences were numbered according to the Chothia scheme [62] using AbNum [53] (www.bioinf.org.uk/abs/abnum/), where missing residues in the numbering scheme sequence were padded with characters dependent on which protein language model was being used, to align all

sequences making V_H sequences 132 residues long and V_L sequences 122 residues long. Details of sequence encodings can be found in Table 5.

5.3 Supervised Machine Learning

Supervised learning was performed with SciKitLearn using 15 classifiers [63] given in Table 6. Descriptions of each classifier used and details can be found in Supplementary File: Supplementary_ML.pdf.

F-regression is a method of feature reduction where the *k* most informative features are kept as input to the model. This is done by calculating the cross-correlation of each data point and the label for all features, which is converted to an F-score, then to a p-value and ranked [64]. F-regression was implemented through the module `sklearn.feature_selection.SelectKBest` using the Python module `sklearn.feature_selection.f_regression` as the score function and variable numbers for *k* were substituted [1,10,50,100,500,1000,2500,5000,10000].

Once the F-regression was implemented on the encoded dataset, it was then split into training and test sets using `sklearn.model_selection.train_test_split` where training portions were used to train the models using ten-fold cross-validation.

Model performance was measured using the Matthews' Correlation Coefficient (MCC) [52], which gives a score between -1 (perfect inverse prediction) and 1 (perfect prediction), with 0 being random chance. Mean MCC and standard deviation for prediction performance over the ten folds were reported.

5.4 Unsupervised Machine Learning

PCA was used as a method of dimensionality reduction and implemented through `sklearn.decomposition.PCA`. Non-linear PCA [46] was implemented through `sklearn.decomposition.KernelPCA` using kernel functions 'rbf', 'cosine' and 'poly' and 2 principal components. At first the coefficient of the kernel (γ) was set to the default value of *1/k* where *k* is the number of features. Once rbf was selected as the most suitable method, differing values for γ were tested [10, 50, 100, 500, 1000]. t-distributed Stochastic Neighbour Embedding (t-SNE) [47] was implemented through `sklearn.manifold.TSNE` with 2 components where the learning rate was set to 10, and the perplexity set to 1000. Uniform manifold approximation and projection (UMAP) [48] was implemented through `sklearn.manifold.UMAP` with the learning rate set to 1 and the nearest neighbours set to 100.

5.5 Ellipse Function

The ellipse function takes in the points of the two extremes on the major axis (*x1, y1*) and (*x2, y2*) as well as a value for *h* (the height of the minor axis). The major axis is taken as the principal

component where clinical mAbs have the largest distribution, and the selected points are given as the points on the distribution closest to a given Z-score in that distribution. The value of *h* is given as the distance between the two equivalent points on the minor axis. The method for producing the ellipse works as follows:

- Calculate the major and minor radii of the ellipse (*a* and *b* respectively). The major radius is calculated from the two given points (Equation 1) and the minor radius is calculated as half the value given for *h*, where *Δx* is the difference in *x* values and *Δy* is the difference in *y* values between the two extreme points on the major axis.

$$a=\frac{\sqrt{\Delta x^2+\Delta y^2}}{2}, b=\frac{h}{2} \quad (1)$$

- Use the parametric equation of an ellipse to generate the ellipse over 100 equally spaced points between 0 and 2π assuming it is centred at the origin (Equation 2). For a given point on the ellipse:

$$x=a\cos(\theta), y=b\sin(\theta) \quad (2)$$

  where *a* is the major axis radius, *b* is the minor axis radius and θ is a given angle between 0 and 2π.

- Calculate the angle between given points to obtain an angle of rotation using the Python Numpy `arctan2` function [65] for *Δy* and *Δx*.

- Calculate a rotation matrix (R) based on the angle of rotation:

$$R=\left[\left[\cos(\theta), -\sin(\theta)\right], \left[\sin(\theta), \cos(\theta)\right]\right] \quad (3)$$

  where θ is the angle of rotation.

- Apply the rotation matrix R, to the ellipse.

- Calculate the midpoint of the two given points:

$$x=\frac{x_1+x_2}{2}, y=\frac{y_1+y_2}{2} \quad (4)$$

- Translate the ellipse to the midpoint.

- For each point, check if its *x* and *y* coordinates are inside the ellipse using the Polygon function from the Python '`Shapely`' module.

5.6 Calculating Physicochemical Properties

Physicochemical properties were calculated as described below and compared between groups using the two-tailed unpaired Mann-Whitney U-test [66].

5.6.1 Identifying CDR-H3 Loops

The Complimentarity Determining Region loop three of the V_H domain (CDR-H3) has frequently been observed to have the largest contribution to antibody binding affinity because it is the most diverse region between sequences, overlapping the Variable, Diversity and Junction gene segments [42, 67]. CDR-H3 regions were identified using the AbNum software [53] and applying the Kabat/Chothia/Martin definition (H95-H102). Sequences with more than two cysteine residues were excluded as additional cysteines are a known risk factor for aggregation [68].

5.6.2 Thermostability

Gibbs Free Energy (ΔG) of unfolding was predicted for each antibody sequence using the Oobatake method [40] with experimental values of ΔH and ΔS taken from the original paper. mAbs with negative ΔG of unfolding values were considered unstable and associated with poor developability. This was calculated for the V_H and V_L chains, as well as for both chains concatenated together using the '`ssbio`' Python module [69].

5.6.3 Isoelectric Point

The method of calculating Isoelectric Point (pI) was that used in the IPC software [41] which uses experimentally obtained peptide pKa values from the EMBOSS database [70] substituted into a rearranged Henderson-Hasselbach equation. The equations are iterated using different pH values, starting at 6.5, and the results of the termini and each of the charged residues are summed together. If the sum is 0±0.01, the isoelectric point is reached. Otherwise, the iteration continues to increase the pH if the summed net charge was positive or to decrease the pH if it was negative.

5.6.4 Immunogenicity (Humanness)

The G score [51], is a measure of antibody humanness based on similarity to germline families and a predictor of immunogenicity. This metric was calculated using the online tool www.bioinf.org.uk/abs/gscore/ for V_H and V_L independently. The minimum score of these chains for each antibody was taken and the mean for each of these sets of minima is presented.

5.6.5 V-region Germline Gene Identification

V-region Germline genes were identified using the in-house 'Assign GermLine' software (AGL; github.com/AndrewCRMartin/agl/). Where more than one germline gene has the same (highest) sequence identity, AGL selects a gene using the logic that the germline family with the lowest family number was likely to have been discovered first and therefore likely to be more numerous. The same logic is applied to allelic variants and proximal genes are favoured over distal genes ensuring that gene names are consistent.

5.7 TAP Scores

TAP scores were developed by Raybould *et al*. [31] to compare an antibody with the clinical dataset using metrics related to developability, assigning 'amber penalties' to antibodies that fall in the top and bottom 5% of the observed distribution, and 'red flags' to antibodies that fall outside the distribution. TAP scores were calculated for 10,382 paired V_H and V_L nucleotide sequences from the Test BCR dataset in batches of 500 using the IGX platform igx.bio/ in August 2023 using the default penalty set. Details of statistics measured and penalties assigned can be found in Raybould *et al*. [31].

References

[1] Rahul Khetan, Robin Curtis, Charlotte M Deane, Johannes T Hadsund, Uddipan Kar, Konrad Krawczyk, Daisuke Kuroda, Sarah A Robinson, Pietro Sormanni, Kouhei Tsumoto, Jim Warwicker, and Andrew C R Martin. Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *mAbs*, 14(1):2020082, 2022.

[2] Mohamed A Alfaleh, Hashem O Alsaab, Ahmad B Mahmoud, Almo-hanad A Alkayyal, Martina L Jones, Stephen M Mahler, and Anwar M Hashem. Phage display derived monoclonal antibodies: From bench to bedside. *Frontiers in Immunology*, 11:1986, 2020.

[3] Matthew I J Raybould, Claire Marks, Alan P Lewis, Jiye Shi, Alexander Bujotzek, Bruck Taddese, and Charlotte M Deane. Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Research*, 48(D1):D383–D388, 2020.

[4] Peter C Taylor, Andrew C Adams, Matthew M Hufford, Inmaculada de la Torre, Kevin Winthrop, and Robert L Gottlieb. Neutralizing monoclonal antibodies for treatment of COVID-19. *Nature Reviews Immunology*, 21(6):382–393, 2021.

[5] Susan J Keam. Tixagevimab + cilgavimab: First approval. *Drugs*, 82:1001–1010, 2022.

[6] Hélène Kaplon, Alicia Chenoweth, Silvia Crescioli, and Janice M Reichert. Antibodies to watch in 2022. *mAbs*, 14(1):2014296, 2022.

[7] Hélène Kaplon, Silvia Crescioli, Alicia Chenoweth, Jyothsna Visweswaraiah, and Janice M Reichert. Antibodies to watch in 2023. *mAbs*, 15(1):2153410, 2023.

[8] Marco Troisi, Eleonora Marini, Valentina Abbiento, Samuele Stazzoni, Emanuele Andreano, and Rino Rappuoli. A new dawn for monoclonal antibodies against antimicrobial resistant bacteria. *Frontiers in Microbiology*, 13:1080059, 2022.

[9] Amy Sun and Leslie Z Benet. Late-stage failures of monoclonal antibody drugs: A retrospective case study analysis. *Pharmacology*, 105:145–163, 2020.

[10] William H Robinson. Sequencing the functional antibody repertoire — diagnostic and therapeutic discovery. *Nature Reviews Rheumatology*, 11(3):171–182, 2015.

[11] Saravanan Rajan, Michael R Kierny, Andrew Mercer, Jincheng Wu, An-drey Tovchigrechko, Herren Wu, William F Dall'Acqua, Xiaodong Xiao, and Partha S Chowdhury. Recombinant human B cell repertoires enable screening for rare, specific, and natively paired antibodies. *Communications Biology*, 1(1):5, 2018.

[12] David B Jaffe, Payam Shahi, Bruce A Adams, Ashley M Chrisman, Peter M Finnegan, Nandhini Raman, Ariel E Royall, FuNien Tsai, Thomas Vollbrecht, Daniel S Reyes, N Lance Hepler, and Wyatt J McDonnell. Functional antibodies exhibit light chain coherence. *Nature*, 611(7935):352–357, 2022.

[13] Sergio E Irac, Megan Sioe Fei Soon, Nicholas Borcherding, and Zewen Kelvin Tuong. Single-cell immune repertoire analysis. *Nature Methods*, 21:777–792, 2024.

[14] Andrew C R Martin. Accessing the Kabat antibody sequence database by computer. *Proteins: Structure, Function, and Bioinformatics*, 25(1):130–133, 1996.

[15] Marie-Paule Lefranc. IMGT, the International ImMunoGeneTics information system. *Cold Spring Harbor Protocols*, 2011(6):595-603, 2011.

[16] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(Database issue):D1140–D1146, 2013.

[17] Saba Ferdous and Andrew C R Martin. AbDb: antibody structure database — a database of PDB-derived antibody structures. *Database*, 2018:bay040, 2018.

[18] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022.

[19] Yicheng Guo, Kevin Chen, Peter D Kwong, Lawrence Shapiro, and Zizhang Sheng. cAb-Rep: A database of curated antibody repertoires for exploring antibody diversity and predicting antibody prevalence. *Frontiers in Immunology*, 10:2365, 2019.

[20] Christian Margreitter, Hui-Chun Lu, Catherine Townsend, Alexander Stewart, Deborah K Dunn-Walters, and Franca Fraternali. BRepertoire: a user-friendly web server for analysing antibody repertoire data. *Nucleic Acids Research*, 46(W1):W264–W270, 2018.

[21] Monica L Fernández-Quintero, Johannes R Loeffler, Johannes Kraml, Ursula Kahler, Anna S Kamenik, and Klaus R Liedl. Characterizing the diversity of the CDR-H3 loop conformational ensembles in relationship to antibody binding properties. *Frontiers in Immunology*, 9:3065, 2019.

[22] Adriana-Michelle Wolf Perez, Nikolai Lorenzen, Michele Vendruscolo, and Pietro Sormanni. Assessment of therapeutic antibody developability by combinations of *in vitro* and *in silico* methods. *Methods in Molecular Biology*, 2313:57–113, 2022.

[23] Olga Obrezanova, Andreas Arnell, Ramón Gómez de la Cuesta, Maud E Berthelot, Thomas R A Gallagher, Jesús Zurdo, and Yvette Stallwood. Aggregation risk prediction for antibodies and its application to biothera-peutic development. *mAbs*, 7(2):352–363, 2015.

[24] Franz Waibl, Monica L Fernández-Quintero, Florian S Wedl, Hubert Ket-tenberger, Guy Georges, and Klaus R Liedl. Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. *Frontiers in Molecular Biosciences*, 9:960194, 2022.

[25] Katharine Bray-French, Katharina Hartman, Guido Steiner, Celine Marban-Doran, Juliana Bessa, Neil Campbell, Meret Martin-Facklam, Kay-Gunnar Stubenrauch, Corinne Solier, Thomas Singer, and Axel Ducret. Managing the impact of immunogenicity in an era of immunotherapy: From bench to bedside. *Journal of Pharmaceutical Sciences*, 110:2575–2584, 2021.

[26] Tushar Jain, Tingwan Sun, Stephanie Durand, Amy Hall, Nga Rewa Houston, Juergen H Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T Gray, Eric M Krauland, Yingda Xu, Maximiliano Vásquez, and K Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114:944–949, 2017.

[27] Tushar Jain, Todd Boland, and Maximiliano Vásquez. Identifying developability risks for clinical progression of antibodies using high-throughput *in vitro* and *in silico* approaches. *mAbs*, 15(1):2200540, 2023.

[28] Timothy M Lauer, Neeraj J Agrawal, Naresh Chennamsetty, Kamal Egodage, Bernhard Helk, and Bernhardt L Trout. Developability index: a rapid *in silico* tool for the screening of antibody aggregation propensity. *Journal of pharmaceutical sciences*, 101(1):102–115, 2012.

[29] Daniel Seeliger. Development of scoring functions for antibody sequence assessment and optimization. *PLOS One*, 8(10):e76909, 2013.

[30] Max Hebditch and James Warwicker. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ*, 7(1):e8199, 2019.

[31] Matthew I J Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116:4025–4030, 2019.

[32] Christopher Negron, Joyce Fang, Michael J McPherson, W Blaine Stine, and Andrew J McCluskey. Separating clinical antibodies from repertoire antibodies, a path to *in silico* developability assessment. *mAbs*, 14(1):2080628, 2022.

[33] Jisun Kim, Matthew McFee, Qiao Fang, Osama Abdin, and Philip M Kim. Computational and artificial intelligence-based methods for antibody development. *Trends in Pharmacological Sciences*, 44:175–189, 2023.

[34] Ganggang Bai, Chuance Sun, Ziang Guo, Yangjing Wang, Xincheng Zeng, Yuhong Su, Qi Zhao, and Buyong Ma. Accelerating antibody discovery and design with artificial intelligence: Recent advances and prospects. *Seminars in Cancer Biology*, 95:13–24, 2023.

[35] Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4(1):23–55, 1985.

[36] William R Atchley, Jieping Zhao, Andrew D Fernandes, and Tanja Drüke. Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences*, 102:6395–6400, 2005.

[37] Ana-Marta Sequeira, Diana Lousa, and Miguel Rocha. ProPythia: A Python package for protein classification based on machine and deep learning. *Neurocomputing*, 484:172–182, 2022.

[38] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, 2021.

[39] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022.07.20.500902, 2022.

[40] Motohisha Oobatake and Tatsuo Ooi. Hydration and heat stability effects on protein unfolding. *Progress in Biophysics and Molecular Biology*, 59(3):237–284, 1993.

[41] Lukasz P Kozlowski. IPC — isoelectric point calculator. *Biology Direct*, 11(1):55, 2016.

[42] Tai Te Wu, George Johnson, and Elvin A Kabat. Length distribution of CDRH3 in antibodies. *Proteins: Structure, Function, and Bioinformatics*, 16:1–7, 1993.

[43] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[44] David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A Bitton. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14:2020203, 2022.

[45] Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, 2023.

[46] Mariëlle Linting, Jacqueline J Meulman, Patrick J F Groenen, and Anita J van der Koojj. Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, 12(3):336–358, 2007.

[47] Matthew C Cieslak, Ann M Castelfranco, Vittoria Roncalli, Petra H Lenz, and Daniel K Hartline. t-Distributed stochastic neighbor embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. *Marine Genomics*, 51:100723, 2020.

[48] Yang Yang, Hongjian Sun, Yu Zhang, Tiefu Zhang, Jialei Gong, Yunbo Wei, Yong-Gang Duan, Minglei Shu, Yuchen Yang, Di Wu, and Di Yu. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Reports*, 36:109442, 2021.

[49] Sofia S Guimaraes Koch, Robin Thorpe, Nana Kawasaki, Marie-Paule Lefranc, Sarel Malan, Andrew C R Martin, Gilles Mignot, Andreas Plückthun, Menico Rizzi, Stephanie Shubat, Karin Weisser, and Raffaella Balocco. International nonproprietary names for monoclonal antibodies: an evolving nomenclature system. *mAbs*, 14:2075078, 2022.

[50] Claire Marks, Alissa M Hummer, Mark Chin, and Charlotte M Deane. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, 37(22):4041–4047, 2021.

[51] Philippe Thullier, Oliver Huish, Thibaut Pelat, and Andrew C R Martin. The humanness of macaque antibody sequences. *Journal of Molecular Biology*, 396:1439–1450, 2010.

[52] Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020.

[53] K R Abhinandan and Andrew C R Martin. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology*, 45:3832–3839, 2008.

[54] Alexander Stewart, Joseph Chi-Fung Ng, Gillian Wallis, Vasiliki Tsioligka, Franca Fraternali, and Deborah K Dunn-Walters. Single-cell transcriptomic analyses define distinct peripheral B cell subsets and discrete development pathways. *Frontiers in Immunology*, 12:602539, 2021.

[55] Andre A R Teixeira, Michael F Erasmus, Sara D'Angelo, Leslie Naranjo, Fortunato Ferrara, Camila Leal-Lopes, Oliver Durrant, Cecile Galmiche, Aleardo Morelli, Anthony Scott-Tucker, and Andrew Bradbury. Drug-like antibodies with high affinity, diversity and developability directly from next-generation antibody libraries. *mAbs*, 13:1980942, 2021.

[56] Paul Arras, Han Byul Yoo, Lukas Pekar, Christian Schröter, Thomas Clarke, Simon Krah, Daniel Klewinghaus, Vanessa Siegmund, Andreas Evers, and Stefan Zielonka. A library approach for the de novo high-throughput isolation of humanized VHH domains with favorable developability properties following camelid immunization. *mAbs*, 15:2261149, 2023.

[57] Marianne Brüggemann, Michael J Osborn, Biao Ma, Jasvinder Hayre, Suzanne Avis, Brian Lundstrom, and Roland Buelow. Human antibody production in transgenic animals. *Archivum immunologiae et therapiae experimentalis*, 63:101–108, 2015.

[58] Xiaobin Xu, Yu Huang, Hao Pan, Rosalynn Molden, Haibo Qiu, Thomas J Daly, and Ning Li. Quantitation and modeling of post-translational modifications in a therapeutic monoclonal antibody from single- and multiple-dose monkey pharmacokinetic studies using mass spectrometry. *PloS one*, 14(10):e0223899, 2019.

[59] Narayan Jayaram, Pallab Bhowmick, and Andrew C R Martin. Germline V_H/V_L pairing in antibodies. *Protein Engineering Design and Selection*, 25:523–530, 2012.

[60] Richard M Goldberg. Lessons learned from the Edrecolomab story: How a checkered past became a checkered flag for monoclonal antibodies in colorectal cancer therapy. *Oncology Research and Treatment*, 28:311–312, 2005.

[61] Alexander Stewart, Emma Sinclair, Joseph C Ng, Joselli Silva O'Hare, Audrey Page, Ilaria Serangeli, Christian Margreitter, Federica Orsenigo, Katherine Longman, Cecile Frampas, Catia Costa, Holly-May Lewis, Nora Kasar, Bryan Wu, David Kipling, Peter J M Openshaw, Christopher Chiu, J Kenneth Baillie, Janet T Scott, Malcolm G Semple, Melanie J Bailey, Franca Fraternali, and Deborah K Dunn-Walters. Pandemic, epidemic, endemic: B cell repertoire analysis reveals unique anti-viral responses to SARS-CoV-2, ebola and respiratory syncytial virus. *Frontiers in Immunology*, 13:807104, 2022.

[62] Cyrus Chothia, Arthur M Lesk, Anna Tramontano, Michael Levitt, Sandra J Smith-Gill, Gillian Air, Steven Sheriff, Eduardo A Padlan, David Davies, William R Tulip, Peter M Colman, Silvia Spinelli, Pedro M Alzari, and Roberto J Poljak. Conformations of immunoglobulin hypervariable regions. *Nature*, 342:877–883, 1989.

[63] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and ´Edouard Duchesnay. Scikit-Learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[64] Xiaowei Yang, Qing Shen, Hongquan Xu, and Steven Shoptaw. Functional regression analysis using an F test for longitudinal data with large numbers of repeated measures. *Statistics in Medicine*, 26(7):1552–1566, 2007.

[65] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler

Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with numpy. *Nature*, 585:357–362, 2020.

[66] Michael P Fay and Michael A Proschan. Wilcoxon-Mann-Whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.

[67] George Johnson and Tai T Wu. Preferred CDRH3 lengths for antibodies with defined specificities. *International Immunology*, 10:1801–1805, 1998.

[68] Stephen R Brych, Yatin R Gokarn, Heather Hultgen, Riki J Stevenson, Rahul Rajan, and Masazumi Matsumura. Characterization of antibody aggregation: Role of buried, unpaired cysteines in particle formation. Journal of Pharmaceutical Sciences, 99:764–781, 2010.

[69] Nathan Mih, Elizabeth Brunk, Ke Chen, Edward Catoiu, Anand Sastry, Erol Kavvas, Jonathan M Monk, Zhen Zhang, and Bernhard O Palsson. ssbio: a python framework for structural systems biology. *Bioinformatics*, 34(12):2155–2157, 2018.

[70] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: the European molecular biology open software suite. *Trends in Genetics,* 16(6):276–277, 2000.

[71] Weining Lin, Jude Wells, Zeyuan Wang, Christine Orengo, and Andrew C. R. Martin. Enhancing missense variant pathogenicity prediction with protein language models using VariPred. *Scientific Reports,* 14: 8136, 2024.

6 Acknowledgements

The test library of paired B cell receptor (BCR) sequences was supplied by Franca Fraternali and Joseph Ng (Institute of Structural and Molecular Biology, UCL). TAP scores were calculated using the IGX platform developed by ENPICOM (NL) with academic licence. This work was supported by the BBSRC LIDo programme under Grant BB/T008709/1.

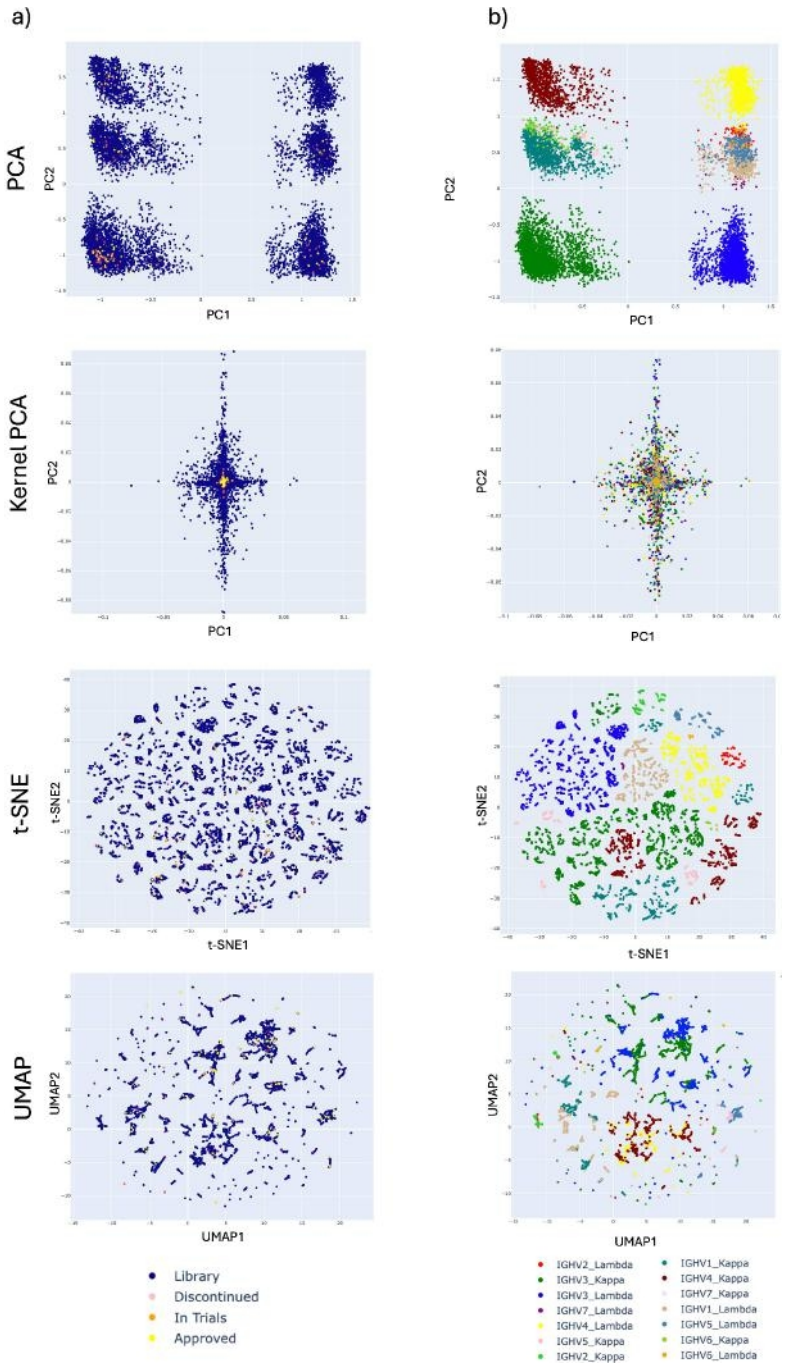

Figure 1 caption: Scatter plots of unsupervised machine learning models trained on clinical (*n*=144) and library (*n*=10,000) paired antibody sequences encoded with the AntiBERTy protein language model. Plots are colour coded by (a) clinical stage or (b) heavy chain V region germline gene and light chain type (λ or κ)

Figure 1 Alt-text: Two columns of scatter plots, from top to bottom: PCA, Kernel PCA, t-SNE and UMAP). The first column is colour-coded to demonstrate where clinical antibodies fit within the human library antibodies and the second column shows the same plots but colour-coded according to heavy chain V-region germline and whether the light chain is a Kappa or Lambda germline. Together, the figures show that, except for Kernel PCA, the heavy and light chain germline pairings have the main influence on the positioning of an antibody in the scatterplot and therefore clinical antibodies are not clustered together, but also follow this positioning. In the case of Kernel PCA, the clinical mAbs are positioned around the origin of the PCA and are radially surrounded by library antibodies.

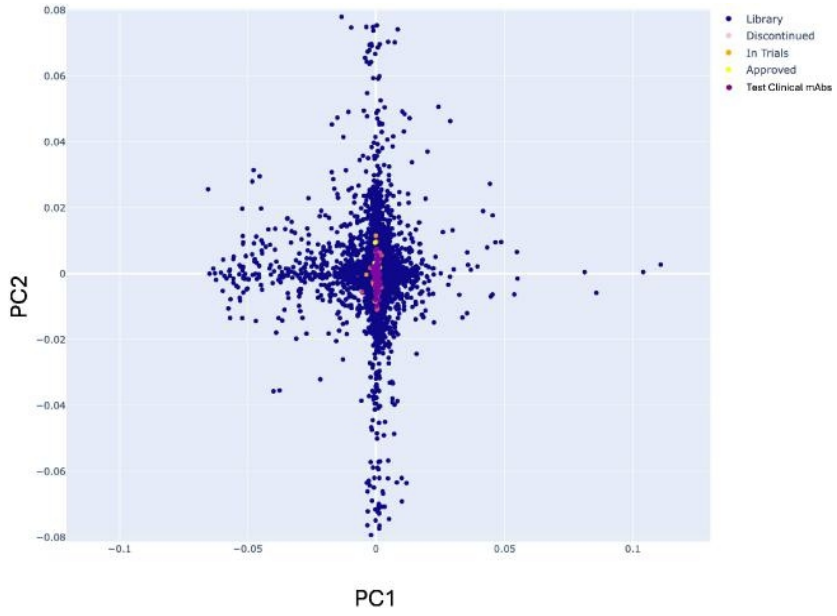

Figure 2 caption: Scatter plot of kernel PCA (kernel='rbf', γ=500) clinical mAbs trained on clinical (*n*=144), library (*n*=10,000) and a held-back test set of clinical (*n*=203) paired human antibody sequences encoded with the AntiBERTy language model.

Figure 2 Alt-text: The scatter plot demonstrates that the held-back test set of more recently named clinical antibodies are also positioned near the centre of the plot, alongside the clinical antibodies demonstrated in Figure 1.

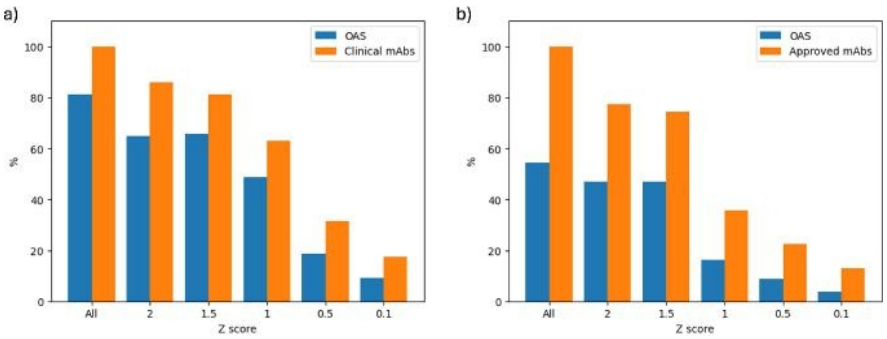

Figure 3 caption: Percentages of OAS (library) and a) human clinical mAbs of any developmental stage, or b) only those with market approval, captured by the ellipse function drawn from the distribution of clinical mAbs. Z-scores denote how wide the distributions for the major axis of ellipse may be drawn with 'All' representing a Z-score selected such that all of the clinical (or approved, respectively) antibodies are captured.

Figure 3 Alt-text: Two bar charts of the proportion of library and clinical antibodies captured by the ellipse function at different Z score cut-offs. The bar charts demonstrate that, as the Z-score threshold is reduced, the proportion of both the OAS antibodies and the human clinical or human approved antibodies retained is reduced. Using all human clinical antibodies, the separation between clinical antibodies and OAS antibodies is relatively constant, although the best separation is seen with a Z-score of 2.0. Using approved antibodies, the 'All' Z-score gives the best separation between approved antibodies and OAS antibodies.

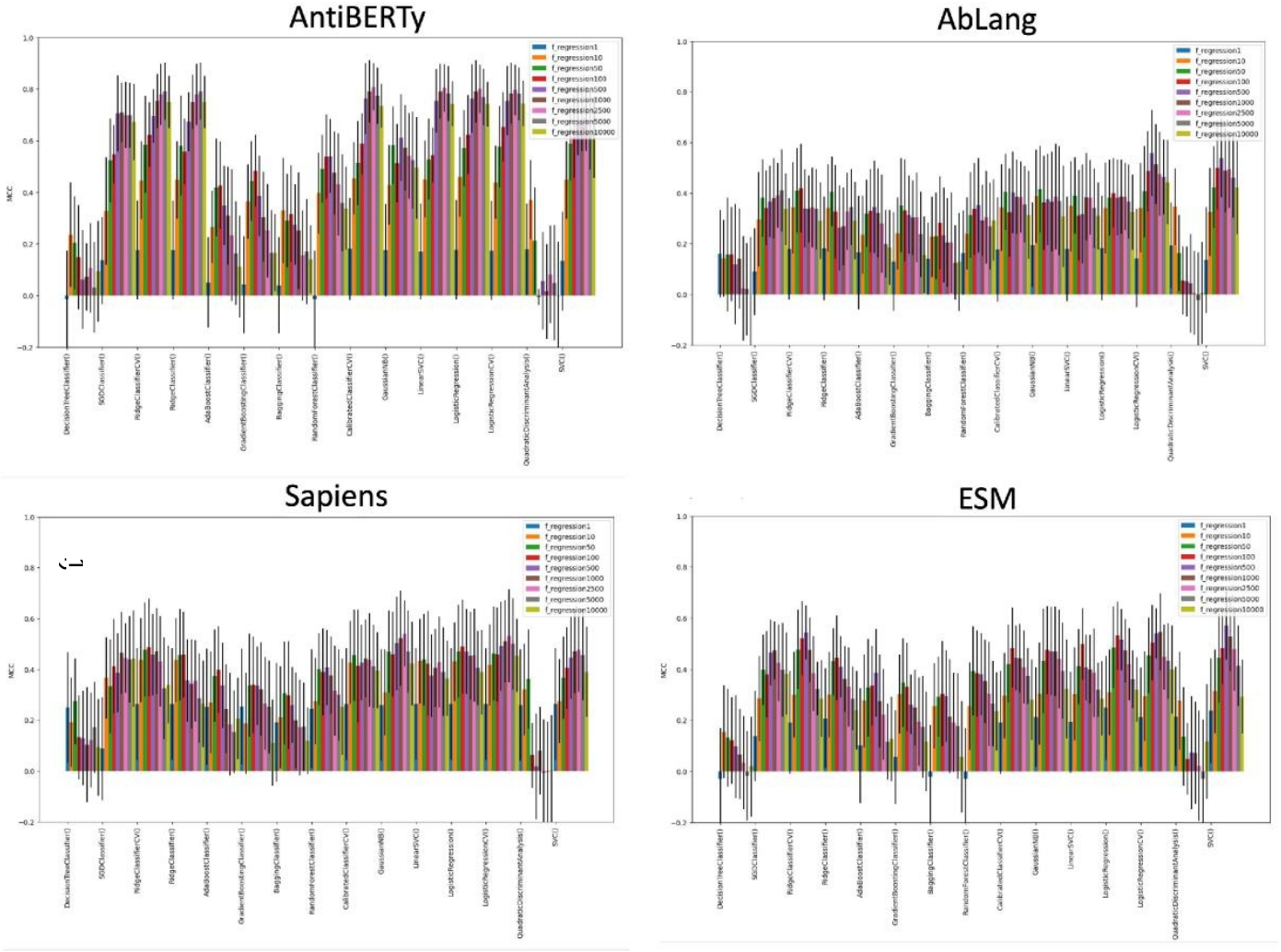

Figure 4 caption: Matthews' Correlation Coefficient (MCC) and standard deviation from ten-fold cross validation of 15 binary machine learning predictors classifying approved (*n*=115) and discontinued (*n*=150) therapeutic antibodies and encoded using four protein language models.

Figure 4 Alt-text: The MCC values and standard deviations are shown on bar charts. Increasing values of *k* are used for the F regression cut off [1, 10, 50, 100, 500, 1000, 2500, 5000, 10000] and results are given for each classifier. The four bar charts show the scores obtained by encoding the training data sequences with different language models (top row: AntiBERTy; AbLang. Bottom row: Sapiens, ESM).

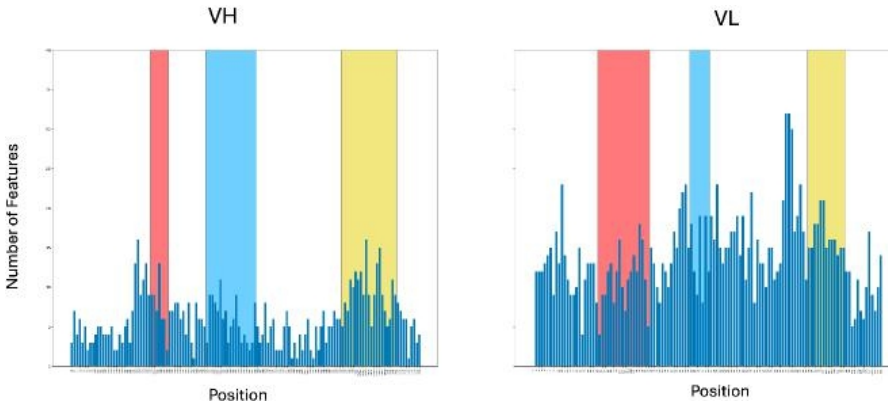

Figure 5 caption: Locations of the top *k* features selected by F-regression from V_H and V_L chains of approved and discontinued mAbs encoded with the AntiBERTy language model [45] where *k*=2500. CDR loops are highlighted in red (CDR1), Blue (CDR2) and Yellow (CDR3).

Figure 5 Alt-text: Two bar charts demonstrating the number of features given for each position across the amino acid sequence aligned using the Chothia numbering scheme for the V_H (left) and V_L (right) sequences selected by the F regression model where *k*=2500. The bar charts demonstrate that the majority of features are in the third framework region of the light chain.

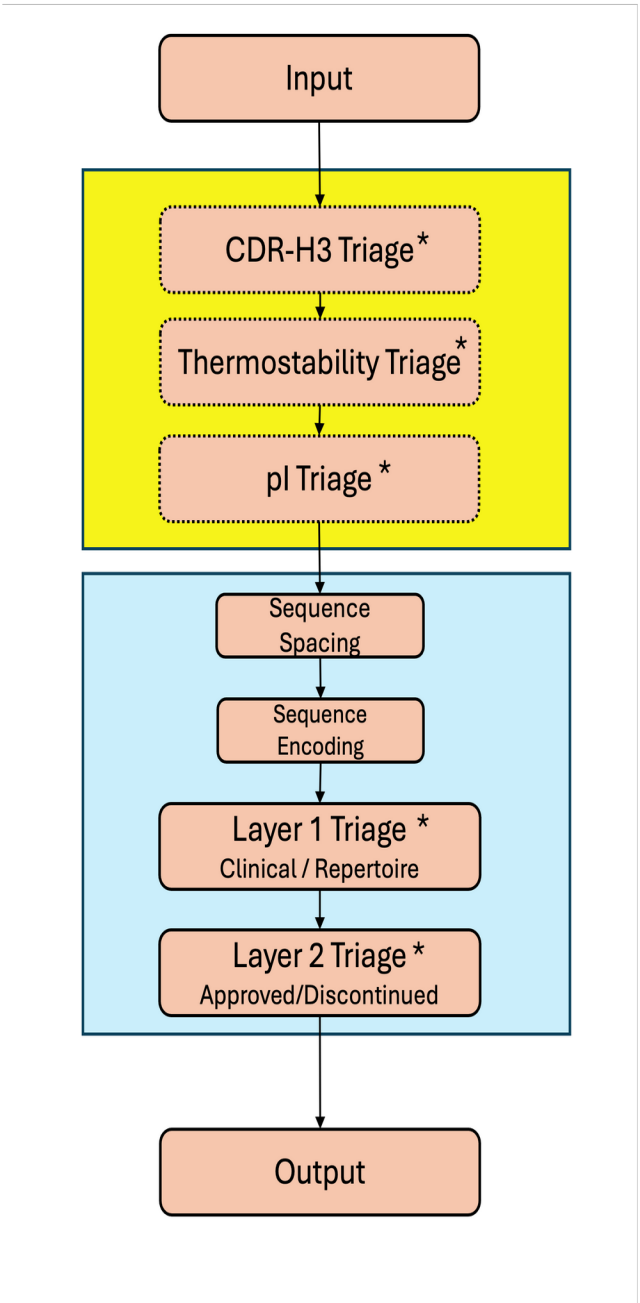

Figure 6 caption: Schematic of the antibody triaging pipeline. The yellow box indicates optional physicochemical feature triaging steps calculating CDR-H3 length using AbNum [53]. Thermostability (ΔG of unfolding) is calculated using the Oobatake Method [40] and pI using the IPC method [41]. The blue box indicates machine learning elements including spacing and encoding, as well as 'Layer 1' triage which is based on the Kernel PCA model for separating antibodies with similar properties to clinical mAbs from the repertoire. The selection of antibodies to take forward is made using the ellipse function. 'Layer 2' is the supervised LinearSVC model trained to distinguish approved and discontinued clinical mAbs. '*' indicates stages where stringency can be adjusted using Z-score thresholds, or the prediction threshold in the case of 'Layer 2'.

Figure 6 Alt-text: Schematic representation of the pipeline using arrows to demonstrate linear path of steps. From the input the arrow enters a large green box with three successive smaller boxes labelled: CDR-H3 Triage*; Thermostability Triage* and pI Triage*. These boxes have hashed borders. The arrow to the next box leaves the green box and enters a large blue box with four smaller boxes labelled: Sequence Spacing; Sequence Encoding; Layer 1 Triage* and Layer 2 Triage*. The arrow then leads outside of the blue box to another box labelled Output.

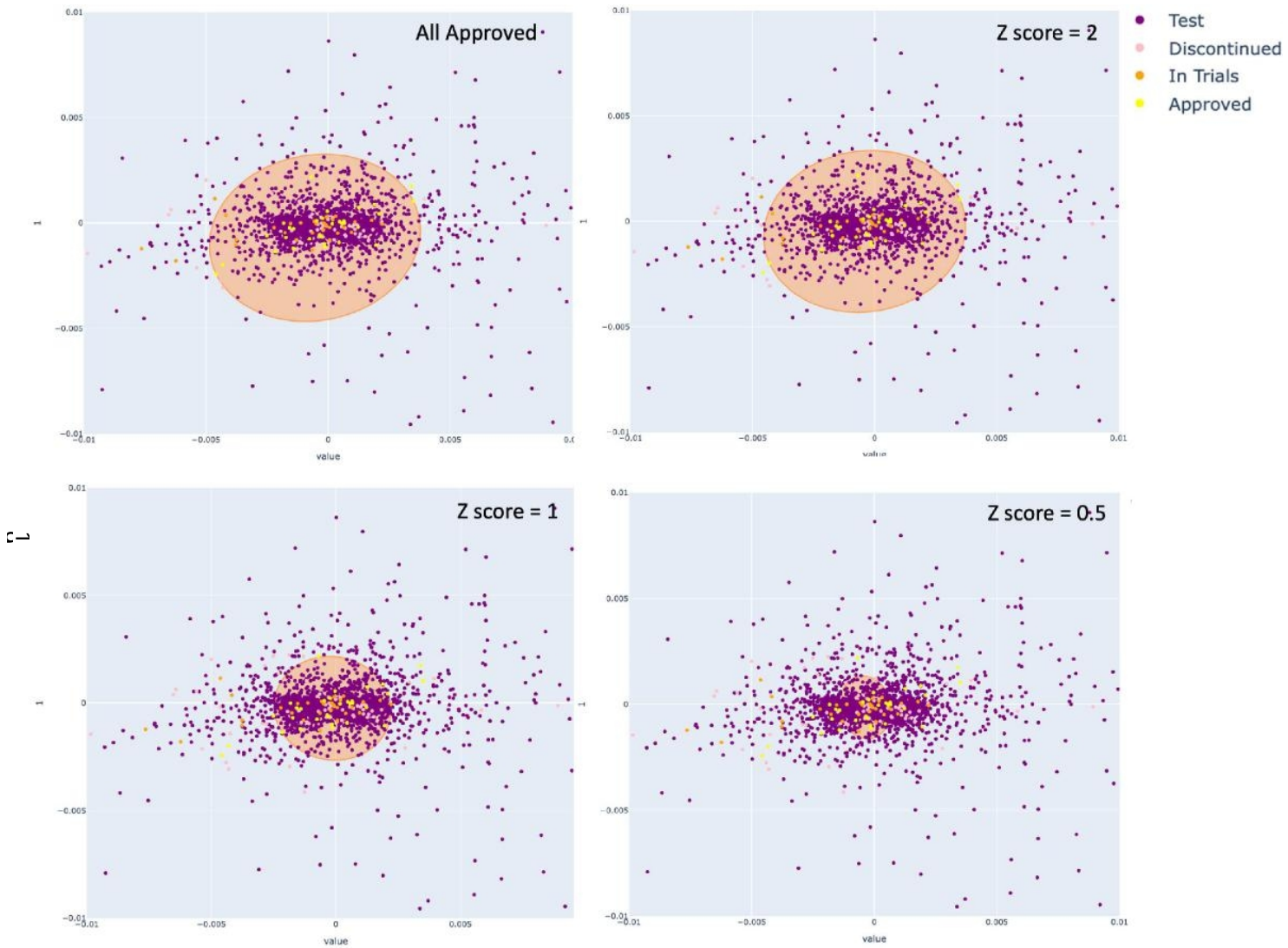

Figure 7 caption: Scatter plots of clinical (*n*=144) and library (*n*=2740) paired antibody sequences encoded with the AntiBERTy protein language model and that have undergone dimensionality reduction using non-linear Principal Component Analysis with a radial basis kernel function (gamma=500). Different Z-scores of the distribution of clinical antibodies along PC1 are used as the extremes of the major axis to draw the ellipse function.

Figure 7 Alt-text: Four scatter plots demonstrating how using Z-scores drawn along the distribution of clinical antibodies (pink, orange, yellow) affects the size of the ellipse drawn to take test antibodies (purple) to the next steps of the pipeline. Each plot shows the size of the ellipse for different Z scores.

Table 1: Means and standard deviation of sequence-calculated physicochemical properties for fully human mAb therapeutics (*n*=144) and repertoire human antibodies from OAS (*n*=10,000, the 'Himan Library Antinodies').

Feature	Human Therapeutic mAbs	Human Library Antibodies	p-value
CDR-H3 Loop Length	12.1±6.65	15.0±10.54	0.00049
ΔG V_H (kJ mol^{-1})	7614±3260	6583±3441	0.00014
ΔG V_L (kJ mol^{-1})	1086±2381	796±2614	0.14
Concatenated V_H/V_L ΔG (kJ mol^{-1})	9248±3896	7944±4238	0.00015
Mean pI of V_H/V_L	7.9±1.30	7.8±1.24	0.025

Table 2: Means of sequence-calculated physicochemical properties for all market approved and discontinued mAbs (including human, humanized, chimeric and murine).

Feature	Approved	Discontinued	p-value
CDR-H3 Loop Length	13.4±4.25	10.7±3.36	0.17
V_H ΔG (kJ mol^{-1})	7008±3806	7592±3424	0.35
V_L ΔG (kJ mol^{-1})	2411±1351	2546±2675	0.33
Concatenated V_H/V_L (kJ mol^{-1})	8434±4855	1071±4094	0.49
Mean pI of V_H/V_L	8.3±1.18	7.9±1.21	0.30
Mean Minimum G score	-1.0±1.22	-0.8±1.06	0.23

Details of the G score are given in Thullier *et al*. [51]

Table 3: Summary performance of the LinearSVC supervised machine learning predictor for success in clinical trials.

	Prediction Threshold	Performance		
		MCC	Sensitivity	Specificity
Cross-validation	0.5	0.80±0.08	0.86±0.10	0.93±0.05
	0.8	0.64±0.11	0.57±0.17	0.99±0.03
Independent	0.5	0.14	0.50	0.64
	0.8	0.51	0.40	1.00

Table 4: Number of antibodies from the Test BCR library output from the triaging pipeline given different parameters of physicochemical filtering and ‘Layer 1’ thresholds. For comparison, the minimum and mean TAP scores are provided, together with the percentage of negative TAP scores, after ‘Layer 1’ and ‘Layer 2’ shown separated by a ‘/’.

		Layer 1 Filtering Z-Score			
		None	2.0	1.0	0.5
None	PCF Onlv	10492	—	—	—
	Layer 1	9875	8165	8186	6107
	Layer 2	3587	2981	2978	2232
	Min TAP Score	-110 / -110	-110 / -110	-110 / -110	-110 / -110
	Mean TAP Score	-18.58 / -18.86	-18.52 / -18.91	-18.53 / -18.97	-18.28 / -18.83
	% TAP Scores < 0	40.3 / 50.4	48.0 / 50.4	30.4 / 50.2	22.2 / 50.7
2.0	PCF Only	8045	—	—	—
	Layer 1	7508	7333	5855	3753
	Layer 2	2571	2514	1981	1272
	Min TAP Score	-110 / -110	-110 / -110	-110 / -110	-110 / -90
	Mean TAP Score	-18.07 / -18.40	-18.05 / -18.47	-18.12 / -18.50	-18.11 / -17.58
	% TAP Scores < 0	44.2 / 47.1	44.1 / 47.2	43.9 / 46.8	19.7 / 47.4
1.0	PCF Only	2740	—	—	—
	Layer 1	2359	2329	2056	1086
	Layer 2	808	797	705	361
	Min TAP Score	-90 / -40	-90 / -40	-90 / -40	-90 / -40
	Mean TAP Score	-16.77 / -57	-16.78 / -16.43	-16.83 / -16.33	-16.99 / -16.23
	% TAP Scores < 0	31.7 / 35.3	31.7 / 35.5	32.3 / 35.6	31.8 / 36.0
0.5	PCF Only	386	—	—	—
	Layer 1	308	231	157	39
	Layer 2	113	80	57	14
	Min TAP Score	-40 / -40	-40 / -40	-30 / -30	-20 / -20
	Mean TAP Score	-15.68 / -15.0	-15.25 / -5.38	-15.33 / -15.00	-11.0 / -12.5
	% TAP Scores < 0	25.3 / 31.9	21.5 / 32.5	28.7 / 38.6	35.6 / 38.6

Physicochemical Filtering (PCF) Z-score

Table 5: Details of language model encodings

Language Model	Features (VH+VL)	Padding Character	Reference
AntiBERTy	130,048	'_'	[45]
AbLang	195,072	'*'	[18]
Sapiens	152,560	'*'	[44]
ESM (esm2 t6 8M UR50D)	82,560	'X'	[39]

Table 6: Supervised machine learning classifiers used in classifying approved and discontinued antibodies.

Classifier	Acronym	Implementation
Decision Tree		sklearn.tree.DecisionTreeClassifier
Stochastic Gradient Descent Classifier	SGDC	sklearn.linear_model.SGDClassifier
Ridge Classifier		sklearn.linear_model.RidgeClassifier
Ridge Classifier CV		sklearn.linear_model.RidgeClassifierCV
AdaBoost Classifier		sklearn.ensemble.GradientBoostingClassifier
Gradient Boost Classifier		sklearn.ensemble.GradientBoostingClassifier
Bagging Classifier		sklearn.ensemble.BaggingClassifier
Random Forest Classifier		sklearn.ensemble.RandomForestClassifier
Calibrated Classifier		sklearn.calibration.CalibratedClassifier
Gaussian Naive Bayes Classifier	GaussianNB	sklearn.naive_bayes.GaussianNB
Support Vector Machine	SVC	sklearn.svm.SVC
Linear Support Vector Machine Classifier	LinearSVC	sklearn.svm.LinearSVC
Logistic Regression Classifier		sklearn.linear_model.LogisticRegression
Logistic Regression CV Classifier		sklearn.linear_model.LogisticRegressionCV
Quadratic Discriminant Analysis Classifier	QDA	sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis