frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 28 February 2024
DOI 10.3389/fimmu.2024.1352703

OPEN ACCESS

EDITED BY
Musie Ghebremichael,
Harvard University, United States

REVIEWED BY
Sai Pooja Mahajan,
Johns Hopkins University, United States
Traian Sulea,
National Research Council Canada (NRC), Canada

*CORRESPONDENCE
Charlotte M. Deane
deane@stats.ox.ac.uk

RECEIVED 08 December 2023
ACCEPTED 30 January 2024
PUBLISHED 28 February 2024

CITATION
Greenshields-Watson A, Abanades B and Deane CM (2024) Investigating the ability of deep learning-based structure prediction to extrapolate and/or enrich the set of antibody CDR canonical forms.
Front. Immunol. 15:1352703.
doi: 10.3389/fimmu.2024.1352703

Investigating the ability of deep learning-based structure prediction to extrapolate and/or enrich the set of antibody CDR canonical forms

Alexander Greenshields-Watson, Brennan Abanades and Charlotte M. Deane*

Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford, United Kingdom

Deep learning models have been shown to accurately predict protein structure from sequence, allowing researchers to explore protein space from the structural viewpoint. In this paper we explore whether "novel" features, such as distinct loop conformations can arise from these predictions despite not being present in the training data. Here we have used ABodyBuilder2, a deep learning antibody structure predictor, to predict the structures of ~1.5M paired antibody sequences. We examined the predicted structures of the canonical CDR loops and found that most of these predictions fall into the already described CDR canonical form structural space. We also found a small number of "new" canonical clusters composed of heterogeneous sequences united by a common sequence motif and loop conformation. Analysis of these novel clusters showed their origins to be either shapes seen in the training data at very low frequency or shapes seen at high frequency but at a shorter sequence length. To evaluate explicitly the ability of ABodyBuilder2 to extrapolate, we retrained several models whilst withholding all antibody structures of a specific CDR loop length or canonical form. These "starved" models showed evidence of generalisation across CDRs of different lengths, but they did not extrapolate to loop conformations which were highly distinct from those present in the training data. However, the models were able to accurately predict a canonical form even if only a very small number of examples of that shape were in the training data. Our results suggest that deep learning protein structure prediction methods are unable to make completely out-of-domain predictions for CDR loops. However, in our analysis we also found that even minimal amounts of data of a structural shape allow the method to recover its original predictive abilities. We have made the ~1.5 M predicted structures used in this study available to download at https://doi.org/10.5281/zenodo.10280181.

KEYWORDS

antibody, canonical forms, structure prediction, complementarity determining regions, deep learning

Introduction

Deep learning has revolutionised the field of structural biology with tools such as AlphaFold2 (AF2) (1), RosettaFold (2) and ESMFold (3) that can accurately predict protein tertiary structure from primary sequence. These tools are all trained on the known protein structure landscape derived from the PDB (4) and have been shown to generalise well to proteins that were not seen during training. Several studies have used these models to enrich the existing protein structure landscape by making extensive predictions from the larger available sequence space. Analysis of these predictions revealed many examples of structures that are very different from the closest available match in experimentally defined data (3, 5).

By analysing over 365,000 high confidence structures predicted by AF2, Bordin et al. were able to define 25 novel superfamilies which did not cluster into any existing CATH classifications using their CATH-Assign protocol (5). A second example of new knowledge arising from structural predictions was provided by ESMFold (3). Here, Lin et al. predicted the structures of over 600M metagenomic sequences isolated from diverse environmental and clinical samples. The use of these metagenomic sequences increased the probability of finding examples that were highly distant from the sequence and structural data used to train ESM2 and ESMFold respectively (3). Within a sample of 1M modelled structures defined as high confidence (predicted local distance difference test score, pLDDT > 0.7 and predicted template modelling score, pTM > 0.7), the authors found over 125,000 predictions with no close match in the PDB [defined as pTM > 0.5 carried out using Foldseek (6)] and in close alignment to the corresponding predictions from AF2. While both studies demonstrate that structure prediction tools can confidently generate novel structures, X-ray crystallography data was not obtained to conclusively validate the predictions. It is also not clear if the novel structures generated are composites of large substructural fragments present in the training data.

To attempt to explicitly address whether models can generalise to unseen regions of structural space, Ahdritz et al. carried out 'out-of-domain' experiments using OpenFold (7). In particular, examining if OpenFold can generalise from limited data to accurately predict alpha helices or beta sheets despite their omission from training datasets. However, they were not able to completely remove all signal of these secondary structures from their training data, and hence the models were likely still learning from a much-reduced set of examples, rather than extrapolating to a completely unknown structure based on their induction of biophysical rules.

These analyses raise the question of whether current deep learning-based models are truly capable of predicting conformations which are never present in training data. While extrapolation by deep neural networks is theoretically plausible (8, 9) searching for evidence of this is difficult and requires extensive classification of training data and the resulting predictions.

One limitation of deep learning based protein structure predictors is their poor performance on stretches of sequence that are intrinsically disordered (10, 11) or explore diverse conformational space (12). The loops of adaptive immune receptors, antibodies, and T cell receptors, fall into the latter category. These loops form the majority of the binding site (paratope) of these proteins and are termed complementarity determining regions (CDRs) (13).

The protein sequences that make up the CDR loops arise from two genetic mechanisms, termed V(D)J recombination (14, 15) and somatic hypermutation (16). In antibodies, the process of V(D)J recombination randomly pairs V-, D- and J-genes (VJ genes for light chains and VDJ genes for heavy chains) and introduces junctional diversity through insertion and deletion of nucleotides. Further diversity is introduced to the V-gene region of the antibody by somatic hypermutation, where point mutations that modify the amino acid sequence and improve binding affinity are positively selected and progressively dominate the immune response to a pathogen. These mechanisms create high levels of sequence diversity and are evolutionarily advantageous as they combine to provide a nearly limitless potential of binding solutions which allow antibodies to neutralise the correspondingly limitless diversity of pathogens to which humans can be exposed (17). Of the six CDR loops on an antibody, diversity is highest within the CDRH3, the residues of which often disproportionally govern paratope-epitope interactions (18). Structure predictors have been found to perform poorly on this region of antibodies, for example with average RMSD values between predictions and ground truth that exceed 2.5 Å for state-of-the-art models such as AlphaFold-Multimer (AFM) (19) and ABodyBuilder2 (ABB2) (20).

The predictive performance on the remaining five loops, CDRL1-3 and CDRH1-2, is far better (average RMSD <1Å), despite these being subject to the genetic process of somatic hypermutation and being influenced by neighbouring hypervariable loops (21). The ability to accurately model these can be explained by canonical forms, the term given to sets of CDR loops of the same length that adopt similar backbone conformations and share a sequence motif (22). These canonical forms were first observed in crystallographic datasets of available antibody structures before 1986 (23). With the deposition of more structural data both the number of canonical conformations and the sequences that could be assigned to each were continuously expanded and redefined (24–29). This information linked diverse sets of sequences to distinct loop conformations and thus was increasingly useful to antibody researchers, by providing a form of sequence-to-structure prediction that could be automated by template search and homology modelling tools (27, 30).

The latest CDR structure and sequence pairings harvested from antibody structural data are defined in PyIgClassify2 (28). The definitions were released as the 'penultimate classification of

Abbreviations: ABB2, ABodyBuilder2; AF2 AlphaFold2; AFM AlphaFold-Multimer; CDR, complementarity determining region; DBSCAN, density-based spatial clustering of applications with noise, DBS, DBSCAN-based selection; DTW, dynamic time warping; KNN, K-nearest neighbours; IMGT, the international ImMunoGeneTics information system; MDS, multidimensional scaling; OAS, observed antibody space; OPIG, Oxford protein informatics group; PDB, protein data bank; RMSD, root mean-squared deviation; SAbDab, the structural antibody database; SCFV, single chain variable fragment.

canonical forms' in reference to the breakthroughs in structure prediction research that may soon render the predictive power of this relationship obsolete. While structure prediction methods are still being evaluated, especially in the domain of adaptive immune receptors, PyIgClassify2 can serve as a map of the known conformational space explored by antibody CDRs.

Using the rigorous definitions from PyIgClassify2 as a reference point in structural space, we set out to test whether the predicted structures from all available paired antibody sequences in observed antibody space, OAS, (31, 32) reveal novel canonical clusters or highlight conformations not explored by existing experimental data. We then assess these new areas to determine whether they represent evidence of extrapolation or had direct origins in the training data.

ABodyBuilder2 (ABB2) is a structure prediction tool specific to antibodies (20). It uses an ensemble of four deep learning models trained on the structures of over 3500 antibodies as well as a fast minimisation in the AMBER14 forcefield (33, 34) to make predictions with comparable accuracy to AFM in a fraction of the time. We used ABB2 to predict the structures of ~1.5M paired antibody sequences. We mapped the conformational space of the CDRL1-3 and CDRH1-2 loops and used existing classifications of the canonical forms in experimental data as reference points. By comparing the loop conformations of canonical clusters to clusters found in predictions derived from the ~1.5M heterogeneous sequences we were able to redefine and identify new canonical clusters.

These new clusters (potential canonical forms) were defined by unique sequence motifs and shared loop conformations and typically arose from enrichment of a small number of examples in the experimental data. We also observed apparently novel clusters (canonical forms) that derived from similar shapes (and sequences) of a different loop length, a phenomenon that has been previously described within the structural dataset (26), termed length independence.

Using our mapping of structural space and the definitions of canonical forms we designed out-of-domain retraining experiments which explicitly tested the capability of ABB2 to both generalise and extrapolate. We found that with zero examples of a given CDR shape ABB2 was consistently unable to predict it. However, with the introduction of very small numbers of a shape, the predictive ability was restored. Overall, these analyses exemplify the power of augmenting experimental data with predictions and provide simple tests for extrapolation and effective data recapitulation that may help inform the next generation of structure predictors.

Methods

Selection of paired antibodies sequences for ABodyBuilder2

Paired antibody sequences were retrieved from observed antibody space (OAS) (31, 32) (1-March-2023, https://opig.stats.ox.ac.uk/webapps/oas/). Non-redundant sequence pairs were filtered by minimum lengths defined in the ABB2 workflow such as minimum residue length of 70, starting IMGT residue number less than 8, and end residue number greater than 120, (see sequence checks https://github.com/oxpig/ImmuneBuilder). Sequences containing any gaps or ambiguities were removed to leave 1,492,044 pairs for processing. ABB2 was run on all sequences and 1,492,031 structures were successfully predicted.

Annotation of CDR loops in paired antibody sequences

To group the relevant modelled structures for conformational analyses, the CDR loops of all input sequences were annotated with information on their sequence composition and length. The CDRs were defined according to the IMGT numbering scheme (CDR1: 27-38, CDR2: 56-65, CDR3: 105-117) (35). This numbering system was chosen as the anchor residues and CDR locations are consistent for both heavy and light chains, as well as being the standard reference point for V(D)J gene annotation. Each sequence was IMGT numbered using ANARCI (36) and the CDR sequences checked against corresponding information from IgBLAST annotations (37). For a given heavy or light chain, if there was any discrepancy between IgBLAST and ANARCI CDR definitions then this chain was not taken forward for conformational analysis (resulting in the exclusion of 6414 light chains and 9822 heavy chains from the structures predicted above).

Retrieval and selection of experimental structures from SAbDab

The structural antibody database (SAbDab) (38, 39) is a curated database which contains all antibody, single chain variable fragment (SCFV) and nanobody structures available in the PDB (4). IMGT numbered structures used in ABB2 test, train and validation datasets were downloaded from SAbDab. These structures were derived by X-ray crystallography or cryogenic electron microscopy (cryo-EM) and with a resolution better than 3.5 Å (full list given in SI of 20).

Annotation of SAbDab structures with PyIgClassify2 information

The information on CDR loop canonical forms was obtained from the pyig_cdr_data.txt file downloaded from the PyIgClassify2 website (http://dunbrack2.fccc.edu/PyIgClassify2/, 21-Feb-2023). This provides complete information for all CDR loops on each structure within a given PDB file, including information on sequence identical but structurally distinct members of the asymmetric unit which are distinguished by their PDB chain identifier. For each loop the relevant information includes the length, sequence, canonical form assignment (both with and without an electron density confidence cut off), PDB identifier of the parent chain and information on whether the CDR is structurally complete or is missing any backbone coordinates. This data was used to filter the relevant structures and their CDR

loops for each analysis, and to annotate the experimental data points by canonical cluster membership.

Alignment of IMGT and AHo numbering systems

PyIgClassify2 CDR lengths are defined according to the AHo numbering scheme (40) which symmetrically places insertions and deletions around positions defined as key residues in each CDR. This deviates from the IMGT numbering scheme (35) which places insertions centrally within each CDR at fixed positions. The different approaches to defining the CDRs mean that IMGT defined lengths for CDRL1-2 and CDRH1-2 are shorter than those defined in the AHo numbering scheme used in PyIgClassify2. Therefore, for each CDR and length combination analysed in this study, we have listed the corresponding AHo CDR lengths in Table 1, along with the PyIgClassify2 defined canonical forms and sequence motifs described in (28).

Pre-processing of SAbDab datasets

SAbDab files included SCFV structures where the heavy and light chain are part of a single continuous sequence, as well as datasets with multiple sequence-identical copies in the asymmetric

TABLE 1 Alignment of IGMT CDR numbering, AHo CDR numbering and PyIgClassify2 Canonical Forms.

CDR	IMGT Length	Aho Length	PyIgClassify2 Defined Canonical Forms (sequence motifs)	
L1	6	11	L1-11-1 (RASQsISsyLA) L1-11-2 (RASQDIsnYLA)	L1-11-3 (gGDniGDKsVH) L1-11-4 (SGDaLpKKYAY)
	7	12	L1-12-1 (RASqSVSSSYLa) L1-12-2 (RASQSVSSNYLA)	
	8	13	L1-13-1 (SGSSSNIGsNYVS) L1-13-2 (TRSSGsIaSNYVq)	L1-13-3 (QSSQSVYNNNNLA)
	9	14	L1-14-1 (RSStGAVTtSNyAN) L1-14-2 (TGTSSDvGgYNYVS)	L1-14-3 (TGSSSNIGAGYDVH)
	11	16	L1-16-1 (RSSQSLVHSNGNTYLe)	
	12	17	L1-17-1 (KSSQSLLySSNqKNYLA)	
L2	3	8	L2-8-1 (YdaSnrAS)	
L3	8	8	L3-8-1 (qQYyNlWT) L3-8-3 (QQYYSSPT)	L3-8-4 (QQYdssPT)
	9	9	L3-9-1 (QqWDSshwv) L3-9-2 (QQyystPYT) L3-9-3 (QsydsSsvv)	L3-9-4 (ALWYSsHWV) L3-9-cis7-1 (QQyYsYPyT) L3-9-cis7-2 (QHFWgTPRT)
	10	10	L3-10-1 (sSYtSSsTwV) L3-10-2 (cSYAGSstwV) L3-10-3 (QvWDSssdVV)	L3-10-cis78-1 (qQrTHwPPLT)
	11	11	L3-11-1 (QaWDSSlsgvV) L3-11-2 (QStDSSGTYwV)	
H1	8	13	H1-13-1 (aASGfTFssYwmH) H1-13-3 (aASGRTFSSYaMG)	H1-13-4 (aaSGGtFsgYYWS) H1-13-5 (AASGRTFSIYaMG)
	9	14	H1-14-1 (TVtGYSITSdYaWN) H1-14-2 (AVSGGSISssYyWS)	
	10	15	H1-15-1 (tFSGFSLSTSGMGVG) H1-15-2 (tvSGDSiSssdyyWg)	
H2	7	9	H2-9-1 (YIYYSGSTY)	
	8	10	H2-10-1 (wInPgNGdTN) H2-10-2 (AISsdGssTY) H2-10-3 (EIyPGsGSTn)	H2-10-4 (gISSGGgYty) H2-10-6 (WINPsGGsTy)
	10	11	H2-12-1 (RTYYRSKWYNd)	

For each CDR, the IMGT lengths used in these analyses are presented alongside the corresponding AHo lengths and PyIgClassify2 defined canonical forms for that CDR. PyIgClassify2 canonical forms are named according to the CDR (e.g. L1) followed by the Aho length (e.g. -6), and the form number (e.g. -1) to form a unique identifier for each form (e.g. L1-11-6). For each canonical form the consensus sequence motif, as defined in (28) is given in brackets. Uppercase letters of the consensus sequence indicate highly conserved amino acids at a given position, while lowercase letters indicate a less conserved amino acid that was still observed in the cluster of loops found in the PyIgClassify2 analyses. Information is only provided for the loops analysed in this study, i.e., those which corresponded to the most dominant non-redundant sequences in OAS, (see Figure 1B).

unit. To ensure all CDR loops could be correctly identified and consistently aligned, the relevant chains in each dataset were isolated, IMGT numbered and then saved as individual files linked to the corresponding PyIgClassify2 meta data. For SCFV structures the continuous sequence was broken and each fragment treated as an individual chain. If a chain within a dataset could not be numbered with ANARCI, or a specific CDR loop was missing residues (as indicated by the PyIgClassify2 'cdr_ordered' flag), they were not included in structural analyses. This processing resulted in 11821 heavy and light chains from 3355 PDB files which were used for further analyses. As ABB2 predicted structures were all correctly numbered and contained only a single copy in the asymmetric unit, they did not require any pre-processing for downstream analyses.

Structural analysis of CDR loops of the same length

To perform structural analysis, CDR loops of predicted structures were grouped according to their CDR type (i.e., CDRL1 or CDRH2), amino acid length and sequence composition (non-redundant sequences only). These were analysed alongside all relevant loop structures from SAbDab, for these experimental data points redundant sequences were included as they may contain alternate conformations of the same sequence.

To provide a consistent frame of reference for each CDR length, a loop template was chosen from the highest resolution PDB structure available, this structure also had to be classified as representative of a PyIgClassify2 defined canonical form ('is_representative' flag) and thus was not likely to be an outlier or exhibit any structural features that set it apart. All CDR loops in the analysis were aligned to this template by superimposition of the alpha carbon atoms of the 10 framework residues either side of the loop (CDR1: 22-26 & 39-43, CDR2: 51-55 & 66-70, CDR3: 100-104 & 118-122). If superimposition resulted in RMSD values greater than 1.5 Å then these were not taken forward for loop comparisons. For predicted structures, the framework regions were highly consistent and less than 5 loops per analysis were eliminated. For experimental data points the number eliminated due to framework misalignments ranged from 0 to a maximum of 31 for CDRL1-Len-6 (out of 2499 chains), with a median number of 3 data points eliminated across all analyses.

The carbon and nitrogen backbone atom coordinates of the aligned loops were extracted and saved (CDR1: 27-38, CDR2: 56-65, CDR3: 105-117). Atom counts were checked and then all pairwise RMSD values calculated. This resulted in an N-by-N pairwise distance matrix of RMSD values including both predicted and experimental datapoints for each CDR loop type at every length. To limit the size of pairwise matrices, loops of predicted structures were analysed in batches of 42,000. Where a CDR and length had more than this number of non-redundant sequences (see Table 2) subsequent batches of a maximum size of 42,000 were run until all relevant loop structures were analysed. All batches were run through the clustering and visualisation pipeline (see sections below) and then inspected to ensure results were consistent across all analyses. For multi-batch CDRs, graphs of the first batch are shown in main figures and graphs of subsequent batches are provided in the Supplementary Information (Supplementary Figure 7).

Analysis of CDRH3 loops

We do not present CDRH3 analyses due to the comparatively poor prediction accuracy in this region by ABB2, and other tools such as AFM (20). This uncertainty meant had we found "novel" canonical forms or observations in the CDRH3 region, we could not have confidence that they reflected real loop conformations.

Furthermore, when we performed CDRH3 clustering on the high frequency shorter sequences (CDRH3 lengths 12, 13, 14, Supplementary Figure 8), there was a lack of high-density clusters. This was likely related to the more evenly distributed occupancy of structural space as seen from the multidimensional scaling plots. When density-based clusters were found, the logo plots were uninformative with no apparent motif present in the middle of the CDRH3, and enrichments localised to the beginning and end of the loops (Supplementary Figure 8). Given the heterogeneity of CDRH3 we felt these results were to be expected and this region did not warrant further exploration for novel canonical forms.

Structural analysis of CDR loops of different length

For later analyses aimed at discovering length independent conformations, we also calculated the distance between loops of different lengths. The normalised dynamic time warping (DTW) scores were used to quantify the relative distance in groups of loops that included both length matched and mismatched pairs. Loops were aligned as described above to a high-resolution template, then raw DTW scores calculated between the coordinates using the 'dtaidistance' library in python (41). Normalisation was applied by squaring the score, dividing by the number of atoms being compared, and taking the square root (this resulted in a DTW score that is equivalent to RMSD when lengths are matched). The squared score was divided by the maximum number of atoms (i.e., for length 9 versus 8, the score was divided by 27 to account for 9 residues, each comprised of two carbon and one nitrogen backbone atoms).

Density based clustering

The pairwise distance matrices of RMSD or DTW values contain information that represents the structural relationships between all loop conformations in an analysis. To identify clusters of loops with similar conformations within this high-dimensional data, we employed density-based clustering, using the density-based spatial clustering of applications with noise (DBSCAN) function from the scikit-learn library (SK learn) (42). The DBSCAN function takes the input distance matrix and two parameters: the minimum number of data points required to form a

TABLE 2 Results of RMSD and DBS cluster analysis in high frequency CDR lengths.

CDR	IMGT Length	Number of Unique Sequences in OAS	New information arising from predictions?	Origin of novel canonical cluster?
L1	6	21,707	No	N/A
	7	10,212	New canonical cluster	Extra density matching existing unassigned exp data
	8	7,596	No	N/A
	9	13,846	New canonical cluster	Extra density matching existing unassigned exp data
	11	6,992	Sub-division of existing cluster	Uneven distribution of density within existing form
	12	10,565	No	N/A
L2	3	2,280	No	N/A
L3	8	15,805	No	N/A
	9	76,087	No	N/A
	10	56,599	New canonical cluster	Density derived from length- independent conformation and existing data
	11	51,277	New canonical cluster	Density derived from length-independent conformation
H1	8	61,617	Sub-division of existing cluster	Uneven distribution of density within existing form
	9	5,655	No	N/A
	10	25,000	Sub-division of existing cluster	Extra density matching existing unassigned exp data
H2	7	27,769	Sub-division of existing cluster	Uneven distribution of density within existing form
	8	99,003	No	N/A
	10	13,114	No	N/A

For each CDR and IMGT length explored in this study, details of the number of non-redundant sequences (and hence loops structurally analysed) and the corresponding results of structural analyses are given. The new information arising from analysis of the predicted structures could be defined as either the identification of a new canonical cluster, or the sub-division of an existing cluster of loops that had previously been defined as belonging to a single canonical form. Further details are given on whether these clusters arose from length independent conformations and/or existing experimental data points classified as unassigned by PyIgClassify2 (28).
N/A, not applicable.

cluster (min points) and the minimum distance between any two points in the same cluster (epsilon). The number and size of clusters identified in each analysis is highly sensitive to these values and must be optimised based on the input data. Therefore, we systematically calculated these values in a consistent manner for all analyses, allowing us to find the maximum number of clusters within high-dimensional space and assess whether any cluster related to a novel canonical form. Min point values were calculated by taking the square root of either the number of loops being compared, or the same number divided by 2. For epsilon values we performed a K-nearest neighbours (KNN) analysis on the distance matrix for values of K ranging from 2 to 5. For each value of K, the elbow point was taken from a scree plot of all KNN distances, and this elbow point value used as epsilon in subsequent DBSCAN analyses. This resulted in four separate DBSCAN analyses (one for each value of K from 2 to 5). To select the most appropriate analysis for cluster inspection and visualisation, we matched the K value used to determine epsilon with the number of clusters identified by DBSCAN. If there was no match, then the next closest match was taken for the highest value of K. While dominant clusters were often evident and easily found using multiple values of epsilon, the impact of optimising these parameters was most apparent when identifying smaller or overlapping clusters. We call the clusters generated in this way DBS clusters (short for DBSCAN-based selection). Code is available at: https://github.com/oxpig/OAS-CanonicalForms

Inspection of density-based cluster structures and sequences

We manually inspected the loops comprising each DBS cluster by visualisation of both structures and sequences. This allowed us to assess the structural difference between each cluster and relate the sequence logos back to the defined sequences of PyIgClassify2 canonical clusters. The aligned 3D loops that were assigned to each DBS cluster were visualised using PyMol (43). We selected random samples of up to 20 loops from the predicted structures, all of which belonged to a specific DBS cluster. Samples were coloured according to cluster membership and viewed in the same frame. These loops were presented in multiple orientations to highlight backbone differences that led to distinct cluster assignment. For sequence logo plots, sequences from the predicted structures which had been assigned to a DBS cluster were plotted in R using the ggseqlogo package (44). Logo plots are shown in the bitwise format (opposed to the proportion format) to maximise identification of

the dominant amino acid enrichments and motifs specific to each cluster.

Multidimensional scaling visualisation

To simplify the complex high-dimensional pairwise distance matrix and allow for easy visualisation, we applied multidimensional scaling (MDS) to create a 2D representation. The axes of these plots are labelled as MDS1 and MDS2 and represent unitless scales that capture the spatial differences between data points. We used the parallelised MDS function from the 'lmds' R package, before processing and plotting the output data using tidyverse packages (45). These plots were then annotated according to the DBS cluster membership of each data point, or the canonical cluster assignment of only the experimentally derived data points.

ABodyBuilder2 out-of-domain experiments

Out-of-domain experiments involved the removal of all data points related to a specific CDR length, or a specific canonical cluster, from both the training and test datasets of ABB2. The model was then trained on this modified dataset from scratch. The criteria for removing data from ABB2 training samples involved dividing the MDS map into quadrants and selecting the quadrant with the most distinct canonical cluster, i.e. clearly separated from other data points. All experimental data points in this quadrant were excluded. To ensure the removal of all relevant data points from the training data, any samples defined as the excluded canonical form by PyIgClassify2 without an electron density cutoff (using the 'cluster_nocutoff' flag) were also eliminated.

Data inclusion experiments

For out-of-domain experiments where small numbers of the excluded canonical form were reintroduced into the training data, we first added the highest-resolution datasets identified as representative of the missing canonical form (determined by the 'is_representative' flag in PyIgClassify2). Subsequently, we progressively reintroduced the next highest-resolution datasets not designated as representative but still belonging to the high-confidence canonical cluster.

Retraining of ABodyBuilder2

The original ABodyBuilder2 model consists of an ensemble of four models each trained independently. To make a prediction the outputs from each model are averaged and the prediction closest to average is selected as the final output. Of the four models, one utilised a 128-dimensional embedding and three utilised a 256-dimensional embedding. Models were trained until no further improvement was seen in the validation loss after 100 epochs.

To facilitate the training of multiple models, for the initial experiments in this study we retrained a single model with a 128-dimensional embedding (not an ensemble). Each model was trained on either a Nvidia GeForce GTX-1080 Ti GPU or Quadro RTX 6000/8000 GPUs for 150-340 epochs for each training stage (see training methods 20), continuing until no further improvement in validation loss was observed after 50 epochs (half the number used to train original ABB2 models). For specific retraining experiments (those used to confirm data inclusion thresholds important for prediction), an ensemble of models was created, each consisting of one model with a 128-dimensional embedding and three models with a 256-dimensional embedding. Each model followed the original ABB2 protocol (training until no improvement after 100 epochs). This process allowed us to carry out a larger number of experimental runs and only build full models when we had identified the data cutoffs that significantly affected prediction accuracy (assessed by RMSD between predictions and experimental data points).

Results

Dominant CDR lengths in paired sequence space are matched by comparable distributions in structural data

We predicted the structures of ~1.5M paired antibody sequences from OAS (31, 32) using ABodyBuilder2 (ABB2) (20), a state of the art deep learning antibody structure predictor. We examined this structural space for evidence of novel canonical forms.

We analysed the length distributions of the CDRL1, CDRL2, CDRL3, CDRH1 and CDRH2 loops in this dataset by both absolute frequency (Figure 1A) and by non-redundant CDR sequence frequency (Figure 1B). This revealed that many loops belonging to a specific length were dominated by a smaller number of unique sequences. For example, CDRL1 IMGT length 6 had a frequency of 753,690 in 1.49M, of which only 21,700 were unique. Therefore, we decided to focus our structural analysis on the CDR loops and length combinations (e.g. CDRL3 loops of length 9, after this point referred to as CDRL3-Len-9) which had the highest number of non-redundant sequences within the dataset (bars marked with an asterisk in Figure 1B, details and numbers given in Table 2).

The length distributions of the experimentally derived SAbDab (38, 39) structures used to develop ABB2 are shown in Figure 1C. Only structures from the train, test and validation datasets which had information on canonical forms detailed in PyIgClassify2 were analysed (28) (38 datasets used in development of ABB2 were not categorised in PyIgClassify2). The CDR loop and length combinations taken forward for further analysis were also enriched in the structural units used to train ABB2 (Figure 1C), with the minimum number of examples seen by ABB2 during training being 268 for CDRH1-Len-9.

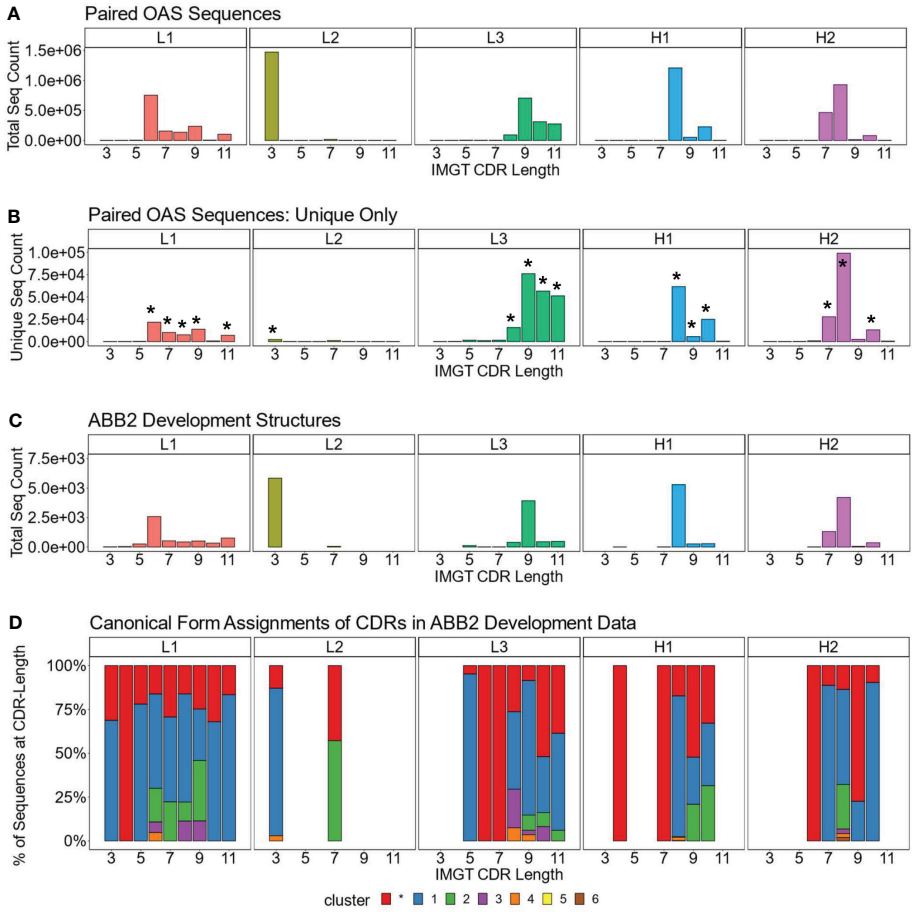

FIGURE 1
Dominant CDR lengths in paired sequence space are matched by comparable distributions in structural data. Frequency distributions of sequences present in OAS by CDR and loop length, for the total number, including all redundant sequences **(A)**, and unique sequences only **(B)**. Asterisks above certain bars in **(B)** indicate the lengths with the most unique sequences, the predicted structures of which were analysed. **(C, D)** show data for the antibodies used to develop ABB2. The loop length frequency for each CDR including all structural units (all copies in the asymmetric unit) which have information in PyIgClassify2 (28) are shown **(C)**. Breakdown of canonical form assignments for corresponding CDR loops present in structures of ABB2 training data **(D)**. Each colour within a bar represents a distinct canonical form, red portions indicate loops that could not be assigned to any canonical form with high confidence in PyIgClassify2 analyses (not labelled with a number in the colour legend).

We next analysed the high confidence PyIgClassify2 canonical form assignments of each CDR loop length marked for further investigation by plotting the proportions of each canonical form within all experimental units (Figure 1D, an experimental unit refers to the fact that one PDB file may have multiple copies in the asymmetric unit). This analysis demonstrated a similar bias, with a single canonical cluster dominating over 50% of assignments for 13 out of the 17 CDR loop and length combinations. Some canonical forms had a very small proportion of examples contained in the ABB2 development data, with the minimum number being 8 examples for CDRH1-Len-8 (Figure 1D). The biases in both the length and canonical clusters distributions of the experimental data indicate that some data-poor areas may benefit from augmentation with predicted structures.

Having selected the CDRs and lengths which dominated our predicted structures, we next built structural clusters for these and explored whether the predicted structures gave rise to new canonical forms or provided insights which were not evident from experimental data alone.

Predicted structures fall into dense regions of conformational space defined by existing canonical forms

We created a map of the structural space for each of the dominant CDR loop and length combinations, identified above, using the predicted antibody structural data. Each map was analysed to find clusters of CDR loops that shared the same backbone conformation. If a SAbDab structure belonged to a cluster this allowed us to annotate the clusters canonical form according to PyIgClassify2. These annotated maps of canonical form structural space enabled us to navigate the predicted structural space and identify highly occupied regions of space not currently defined by a canonical form.

A full description of how these structural space maps were generated is given in the methods. In brief, each map is a 2D representation of the 3D clustered space of a CDR type at a given length. Data points representing loops from both experimental and predicted structures are coloured by their DBS cluster membership.

All data points which are not assigned to a DBS cluster are coloured black. For canonical form annotation the experimental data points are coloured according to their PyIgClassify2 high confidence canonical cluster assignment. Any loops that do not belong to a high confidence PyIgClassify2 cluster (defined by an asterisk in the canonical cluster label) are coloured in red.

Figures 2A, B show the structural space map for CDRL1-Len-6. The projections are overlaid with either DBS cluster membership information (Figure 2A), or canonical cluster classifications from PyIgClassify2 (Figure 2B). The sequences of the loops which comprised these clusters are visualised using logo plots (Figure 2C) to identify the motifs and amino acid enrichments which should match to the canonical sequence motifs described in PyIgClassify2 (Table 1). Samples of loops were also inspected in 3D to assess differences in backbone conformations that give rise to the distinct clusters (Figure 2D).

For many of the CDR and length combinations analysed in this way, the clusters arising from predicted structures aligned well with the dominant clusters of experimentally defined canonical forms and did not highlight any new areas of density without canonical cluster assignment (Figure 2; Supplementary Figure 1; Table 2). Inspection of loop alignments revealed that our RMSD and DBS analysis method could distinguish between backbone kinks, peptide flips and minor variations that equated to less than 1 Å RMSD between data points (Supplementary Figures 2A–C). These minor differences in conformation were also detected by the dihedral angle metric used to compile PyIgClassify2 clusters resulting in similar global divisions of structural space. Before exploring areas that were not accounted for by existing definitions of canonical forms, we next inspected any inconsistencies between our RMSD/DBS analysis and PyIgClassify2.

Differences between PyIgClassify2 definitions and density based structural clusters

While most DBS clusters detected in our analysis could be mapped to experimental data points that adhered to a high confidence canonical form, several of the more subtle PyIgClassify2 definitions were assimilated into a single DBS cluster. We investigated these assimilated data points to assess whether our method was missing important conformational differences.

The loop which best exemplified this was CDRL3-Len-8 (Supplementary Figure 1A). Our analysis pipeline identified two DBS clusters, the centroids of which were 1.45 Å apart and had distinct sequence motifs (Supplementary Figure 1A). Inspection of the PyIgClassify2 canonical clusters demonstrated that cluster 1

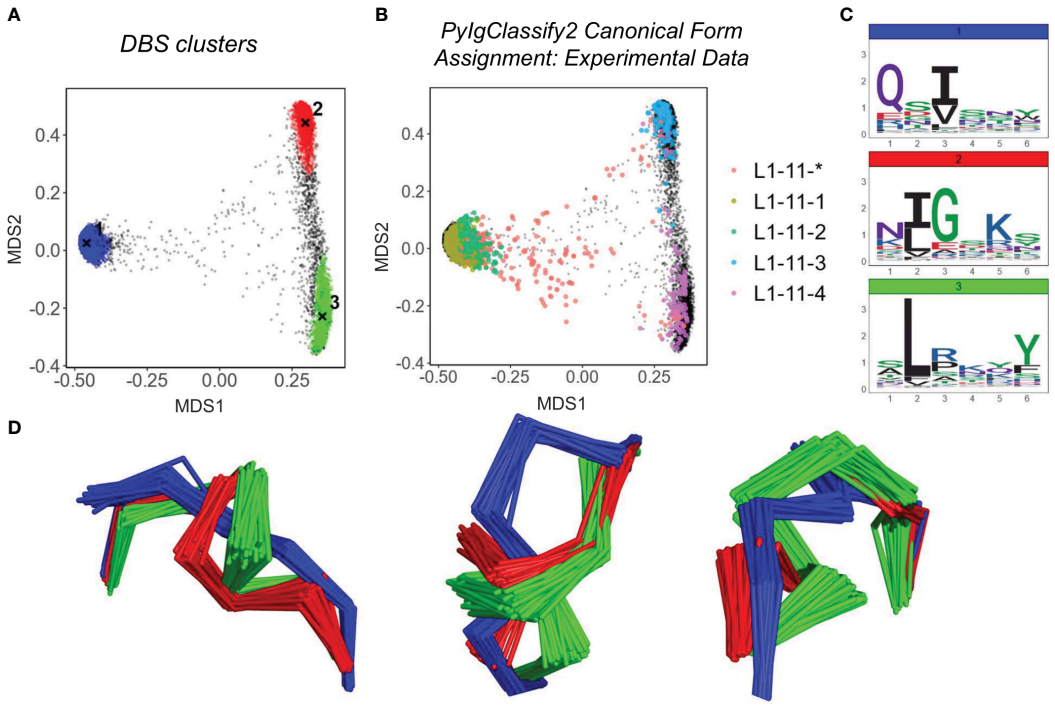

FIGURE 2
Predicted Structures fall into dense regions of conformational space defined by existing canonical classes. Analysis of CDRL1 length 6 loops. Multi-dimensional scaling plots derived from pairwise RMSD data **(A, B)**. Data points in **(A)** are coloured according to density-based clustering (DBS) membership, with the cluster centroid marked by an **X**. Those data points which do not belong to a DBS cluster in **(A)** are coloured black. In **(B)** all experimental data points present in the MDS analysis are coloured according to their high confidence PyIgClassify2 canonical form annotation. Any loops that do not belong to a high confidence PyIgClassify2 cluster (defined by an asterisk in the label, e.g. L1-11-*) are coloured in red. The predicted data points are coloured in black and underly the experimental annotations. Logo plots are shown for all sequences in each DBS cluster **(C)**. For a sample of 20 loops in each cluster, the framework aligned backbone conformations are shown in three different orientations **(D)**. The logo plots **(C)** and backbone **(D)** colours of blue, green and red correspond to the numbered DBS cluster colours in **(A)**.

(coloured blue in Supplementary Figure 1A) defined by the logo motif of QQYysxxT was subdivided into two canonical clusters, of which one had a proline at position 7 (see Table 1, canonical forms L3-8-1 and L3-8-3). We reran our DBS clustering method using an alternate min points term (square root of N data points, opposed to square root of N/2) and found this was able to subdivide the major cluster into two distinct clusters (Supplementary Figure 2E) distinguished by the proline at position 7 (Supplementary Figure 2F). These two clusters are only 0.4 Å apart and exhibited a large degree of overlap in conformational space (Supplementary Figure 2G).

We found additional examples of this within the clusters of loops for CDRL1-Len-6 (Figure 2B) and CDRH2-Len-8 (Supplementary Figure 1C) where the effect was apparent from the smaller number of DBS clusters annotated by a larger number of canonical cluster labels. However, these were often canonical forms assigned to a comparatively low proportion of experimental data points, with more dominant canonical clusters showing greater overlap with our analyses (see later figures). We reasoned that such subtle shifts in conformation would not serve as strong evidence of extrapolation, despite being valid definitions of canonical forms from the dihedral angle perspective. Therefore, their detection was not crucial to our exploratory search for new knowledge arising from structure prediction tools.

Predicted structures enrich the experimental landscape revealing subdivisions of existing classes with defined sequence motifs

Having confirmed that our clustering pipeline was able to pick out major differences in loop conformations across large datasets, we next investigated the structural clusters and sequence logos which did not sit within with PyIgClassify2 defined canonical forms. These ambiguous clusters could be divided into two categories, those which contained experimental data points defined by a canonical form but could be further subdivided into new clusters with distinct sequence motifs and loop conformations, and those which contained experimental data points that were not assigned to a canonical form.

Firstly, the enrichment of the existing structural space with predicted structures led to subdivisions of existing canonical clusters. For example, in CDRH1-Len-8 (Figures 3A, B) and CDRH1-Len-10 (Figures 3C, D) as well as CDRL1-Len-11 and CDRH2-Len-7 (Supplementary Figures 3A, B), we observed DBS clusters with distinct amino acid motifs and loop conformations. These were resolved from the increased structural dataset. The four subclusters observed in CDRH1-Len-8 (Figure 3A) and derived from the PyIgClassify2 canonical form H1-13-5 (Figure 3B, for motifs and length comparisons see Table 1) showed different amino acid patterns at position 2, 4 and 5 of the loops. These subclusters had conformations which differed by RMSD of between 0.79-1.43 Å for cluster centroids. Meanwhile the two subclusters identified from the canonical form H1-15-2 (Figure 3D; Table 1) in CDRH1-Len-10 loops exhibited a difference of 1.09 Å and had sequence patterns that differed at six of the ten positions (Figure 3C).

For all four CDR loops where predictions gave rise to sub clusters, the RMSD between new cluster centroids were often close to the mean value of each analysis (range of mean values from all pairwise comparisons: 0.96 – 1.03 Å, range of distances between cluster centroids: 0.27 – 1.64 Å) and originated from examples present in the training data. Hence the novel canonical classes arising here did not come from generalisation or extrapolation, but simply by statistical power - the increased sensitivity of density-based clustering on the far larger structural set (number of experimental data units versus predicted structures for CDRL1-Len-11: 768 vs 6,992, CDRH1-Len-8: 5,196 vs 61,617, CDRH1-Len-10: 314 vs 2,500 and CDRH2-Len-7: 1298 vs 27,769).

Enrichment of unassigned areas of structural space defines new canonical forms within heterogeneous sequences

The second set of novel clusters identified related to dense areas of predicted structural space where a smaller number of experimental structures existed but were defined as “unassigned” to any canonical form in PyIgClassify2. For example, for CDRL1-Len-7 a cluster made up of 909 predicted data points (and 32 experimental data points) was identified which was distinct from the centroid of two existing canonical clusters by RMSD values of 3.94 and 4.31 Å respectively (Figure 4A). This “new” canonical form has a sequence motif with a strong preference for SGH at positions 1-3 of the loop, in contrast to QSV in both existing forms (see logo plot in Figure 4A, PyIgClassify2 annotations in Figure 4B and corresponding motifs in Table 1). A second example, CDRL1-Len-9 (Figure 4C), had fewer experimental structures within that area (13 were present in training data) and comprised of 673 predictions. The central motif of INV at positions 3-5 showed no overlap with the enriched residues at the same positions within the three existing canonical forms (Figure 4D, see Table 1), with RMSD values between the corresponding cluster centroids of 2.99-3.51 Å. Both observations were enabled by increased population of structural space with ABB2 predictions from heterogeneous sequences, however they are not evidence of extrapolation given their origins in the training data.

New forms exemplify length independent canonical classes arising from somatic hypermutation

Within the CDRL3 loops we found two further examples of highly populated DBS clusters that did not fit with any PyIgClassify2 definitions. These clusters were identified in the analyses of CDRL3 loops of lengths 10 and 11. Both areas of density contained experimental structures that were classified as unassigned to any high confidence canonical cluster by PyIgClassify2 (CDRL3-10 Figures 5A,B, and CDRL3-11 Figures 5D, E).

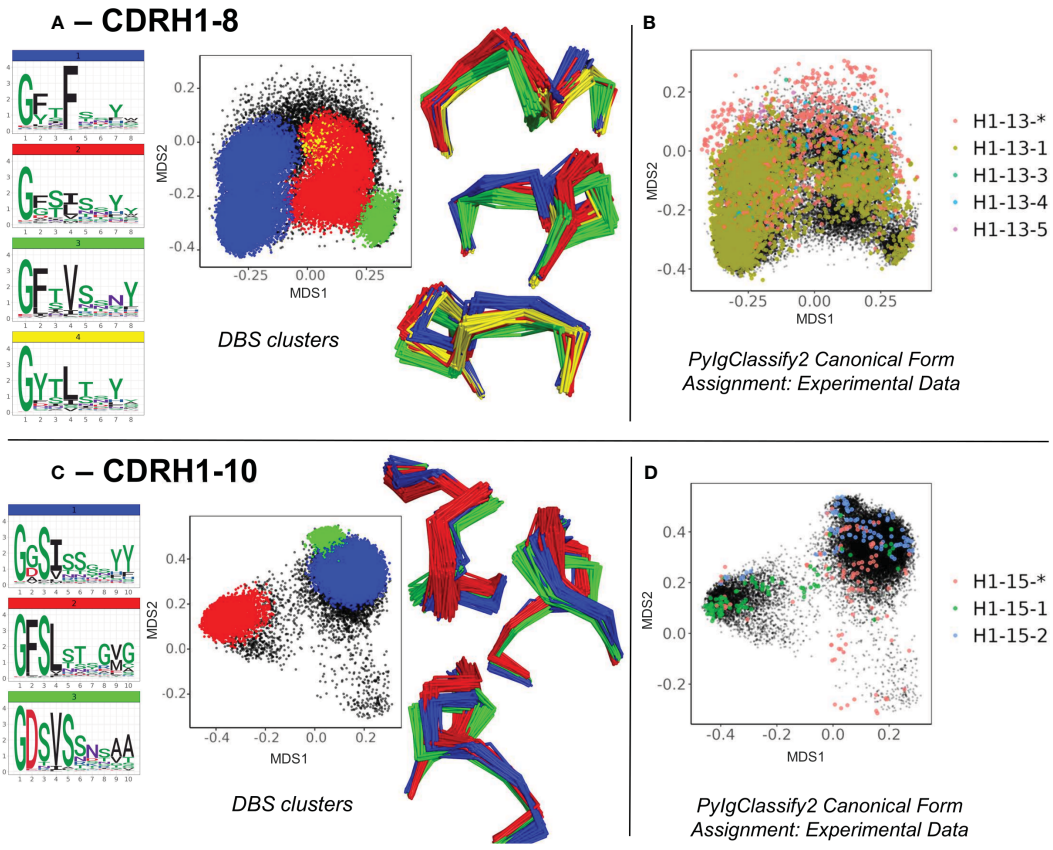

FIGURE 3
Subdivisions of existing canonical classes into clusters with distinct sequence motifs and backbone conformations. For CDR loops of CDRH1 length 8 **(A, B)** and CDRH1 length 10 **(C, D)**, distinct DBS clusters contained multiple experimental data points sharing the same canonical form. **(A, C)** show colour coded DBS cluster logo plots, MDS plots coloured by DBS cluster membership and framework aligned backbone conformations of a sample of 20 loops from each cluster in three different orientations. Data points in the MDS plots which do not belong to a DBS cluster are coloured black. **(B, D)** show all experimental data points present in the MDS analysis coloured according to their high confidence PyIgClassify2 canonical form annotation. Any loops that do not belong to a high confidence PyIgClassify2 cluster (defined by an asterisk in the label, e.g., H1-13-* and H-15-*) are coloured in red. The predicted data points are coloured in black and underly the experimental annotations.

However, these loop shapes are potentially derived from somatic hypermutation (SHM) insertions into CDR loop sequences classified as canonical forms at a shorter length (Supplementary Figures 4A–C). These were evident from inspection of logo plots of CDRL3-Len-10 cluster 3 position 8 (Figure 5C), and CDRL3-Len-11 clusters 1 and 5 at position 9 (Figure 5F), where the motif was nearly identical to that of a highly populated cluster in the CDR one amino acid length below (PyIgClassify2 canonical forms of L3-9-2 and L-10-cis78-1, see Table 1). The SHM insertions were clearly visible on each logo plot as they resulted in no consensus amino acid enrichment at a fixed position in the loop (positions are marked by an arrow in Figures 5C, F respectively).

Given we could identify the corresponding canonical cluster at the shorter length (we termed this the 'origin cluster') (Supplementary Figure 4), we decided to quantify and compare the conformation differences between the two sets of loops. To obtain distance scores that could be used for comparison, both for loops of the same length and differing lengths, we substituted pairwise RMSD calculations with dynamic time warping (DTW) calculations which can be performed on coordinate arrays of differing dimensions (see methods for details).

We represented the structural relationships between CDRL3 loops of length 9 and 10 (Supplementary Figure 5A), and CDRL3 loops of length 10 and 11 (Supplementary Figure 5B) using DTW. The cluster of loops representing the novel conformation in CDRL3-10 (cluster 4: QQYxxxPxxT, coloured yellow in Supplementary Figure 5A) was closest in 3D space (DTW distance between cluster centroids of 0.68 Å) to a cluster composed of CDRL3-Len-9 loops (cluster 1: QQYysxxxT, coloured blue in Supplementary Figure 5A), and only 1.21 Å away from the proposed origin cluster (CDRL3-Len-9 cluster 2: QQyxxxPxT, coloured red in Supplementary Figure 5A). These distances were less than, or comparable to both clusters present at the same length of 10 (distances of 1.83 Å and 1.12 Å apart respectively).

The same effect was more pronounced for the novel conformation in CDRL3-Len-11 (cluster 5: QQYxxxPPxxT, coloured pink in Supplementary Figure 5B). Here the centroids of clusters found at the same length were 2.48 Å and 2.94 Å away,

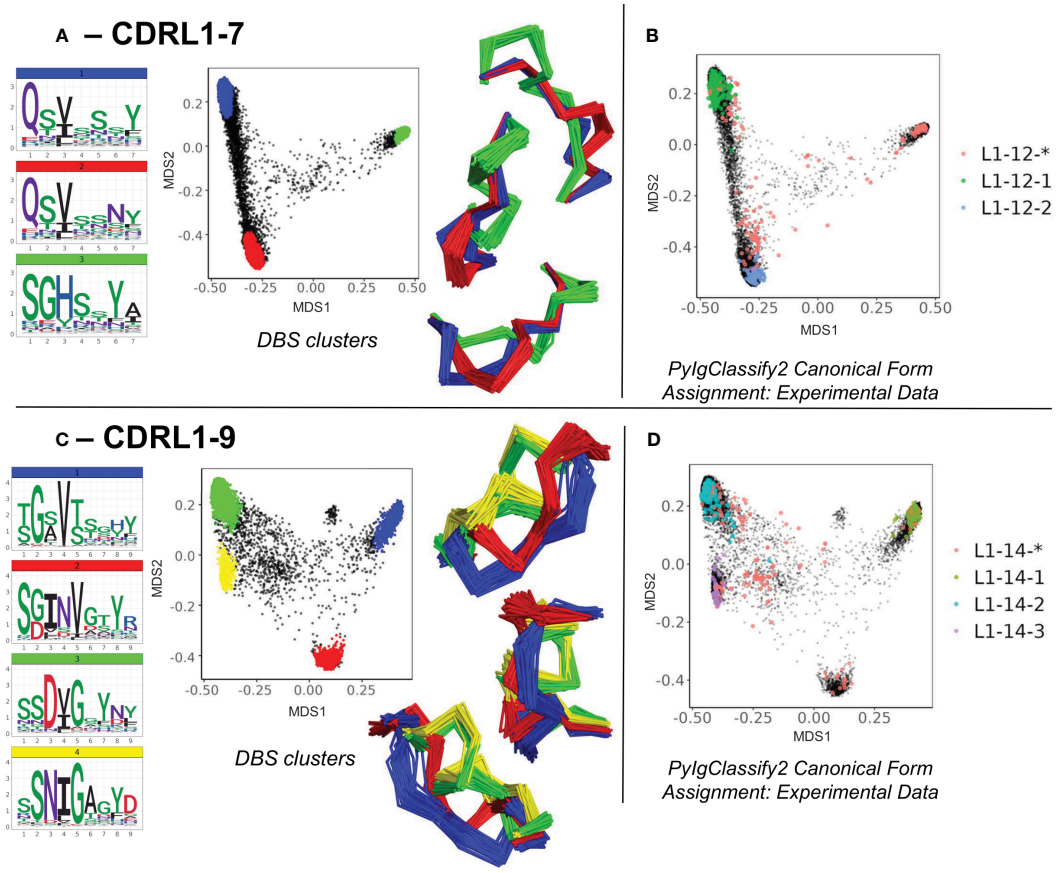

FIGURE 4
Novel canonical classes revealed by structure prediction. Breakdown of analyses where regions of structural space populated by *unassigned experimental data (denoted by red data points with an asterisk in the PyIgClassify2 annotations) resolved to DBS clusters with distinct sequence logo plots for CDRL1 length 7 **(A, B)** and CDRL1 length 9 **(C, D)**. **(A, C)** show colour coded DBS cluster logo plots, MDS plots coloured by DBS cluster membership and framework aligned backbone conformations of a sample of 20 loops from each cluster in three different orientations. Data points in the MDS plots which do not belong to a DBS cluster are coloured black. **(B, D)** show all experimental data points present in the MDS analysis coloured according to their high confidence PyIgClassify2 canonical form annotation. Any loops that do not belong to a high confidence PyIgClassify2 cluster (defined by an asterisk in the label, e.g., L1-12-* and L-14-*) are coloured in red. The predicted data points are coloured in black and underly the experimental annotations.

while the cluster at the shorter length was only 2.07 Å away. Visual inspection of the loop backbones from different clusters shows how similar conformations of mismatched lengths (Figures 5H, J) can be closer in 3D space than matched lengths from different DBS clusters (Figures 5G, I).

These observations fit with the previously described idea of length independence in canonical forms (26), where the closest partner of a structural cluster, or CDR loop, is another cluster, or loop, present at a different length. We hypothesise that the high frequency of predicted structures derived from heterogenous sequences in OAS altered by SHM, helped to reveal these length independent patterns.

ABB2 can generalise across CDRL3 loops which differ in length by one amino acid

To explicitly test whether the training and test data points with a specific CDR length influence the predictions of CDR loops of a different length we performed several out-of-domain experiments. These involved modifying the ABB2 training and test data to remove all data points containing CDRL3 loops length of 8, 9 or 10 (one length per experiment). In each case a new instance of the ABB2 model was trained on a reduced dataset. The resulting models were used to make predictions from sequences of the withheld length which could be assessed individually for accuracy, and together for occupancy of structural space.

The first model tested was trained in the absence of all 392 datapoints (referring to all copies in the asymmetric unit of each PDB file) that had CDRL3 length 8 loops. Prediction accuracy was poor (Supplementary Figures 6A, B), with median RMSD values (prediction versus ground truth) of 1.41 Å compared to 0.46 Å for the fully trained ABB2 model (Figure 5K). There was no clear separation of canonical clusters in conformational space, with median values for each form above 1 Å from the ground truth structure (Supplementary Figure 6B).

In contrast the model trained in the absence of CDRL3-Len-9 data had better prediction accuracy on the withheld structures

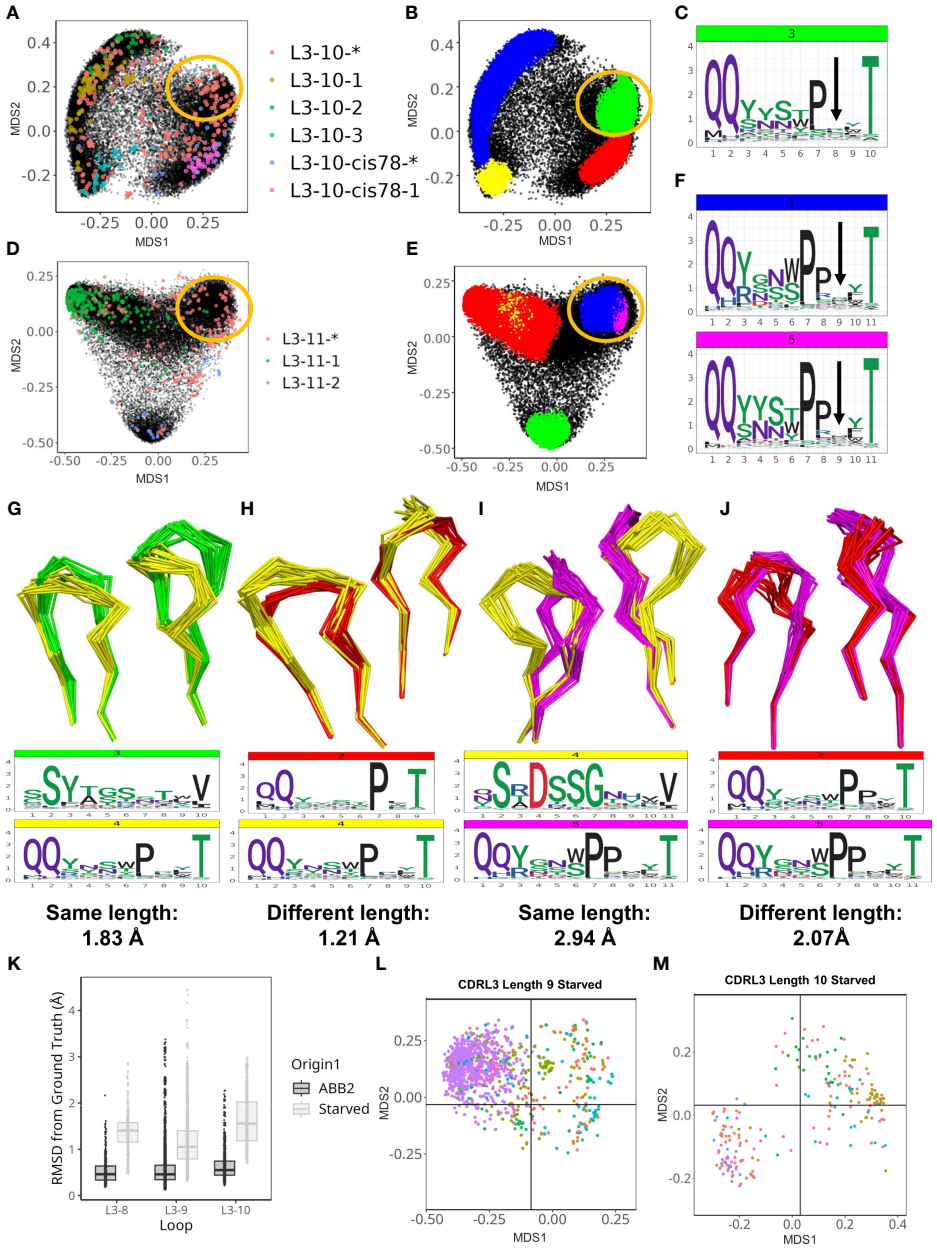

FIGURE 5
ABB2 can generalise across CDRL3 loops which differ in length by one amino acid. Novel DBS clusters were identified for CDRL3 length 10 **(A–C)** and CDRL3 length 11 **(D–F)**. **(A, D)** show experimental data points coloured according to their high confidence PyIgClassify2 canonical form annotation. Any loops that do not belong to a high confidence PyIgClassify2 cluster (defined by an asterisk in the label, e.g. L3-10-* and L3-11-*) are coloured in red. The predicted data points are coloured in black and underly the experimental annotations. Data points in **(B, E)** are coloured according to density-based clustering (DBS) membership, with the data points which do not belong to a DBS cluster coloured black. The areas circled in yellow on all four MDS plots **(A, B, D, E)** relate to the DBS clusters of predicted data points, containing experimental data not assigned to any canonical form that were likely to have arisen from length independence. Logo plots were shown with an arrow indicating the high entropy position with no consistent enrichment **(C, F)** and likely somatic insertion into a shorter canonical cluster (see Supplementary Figure 5). Dynamic time warping analysis permitted quantification of cluster distances between CDR loops of different length as well as clusters of the same length. Clusters of the same length for CDRL3 length 10 are compared by visualising the backbone atoms and sequence logo plots **(G)**, while the novel cluster (coloured yellow) is compared against the proposed origin cluster that the short length of 9 in **(H)**. The DTW distance between cluster centroids is given below each logo plot. CDRL11 clusters of the same length are compared in **(I)**, then the novel cluster and proposed origin cluster in CDRL3 length 10 are compared in **(J)**. Out-of-domain experiments were carried out by retraining ABB2 in the absence of all experimental data points for each of the CDRL3 lengths 8, 9 and 10 [**(K–M)**, also see Supplementary Figure 6]. Boxplots (median and upper and lower quartiles) and dot plots of RMSD values between predictions and ground truth for each starved model versus the original ABB2 ensemble are compared **(K)**, each dot corresponds to the RMSD value of one comparison. The global conformational space of predictions on withheld data points specific to each model are shown for the CDRL3 length 9 **(L)** and CDRL3 length 10 **(M)** starved models. The separation of data points according to canonical form classification was compared to true conformational space and ABB2 ensemble predictions in Supplementary Figure 6 (for each length MDS calculation was performed all data points in the same analysis to allow comparison).

(3865 datapoints, median RMSD 1.05 Å) but still worse than the fully trained ABB2 ensemble (median RMSD 0.46 Å). However, the MDS representation of conformational space showed early separation of data points defined by similar canonical clusters (Figure 5L; Supplementary Figures 6C, D), indicating some rationalisation of the sequence to structure relationship via length offset data. Accuracy for the CDRL3-Len-10 model was the worst (median RMSD 1.66 Å Figure 5K), however a higher standard deviation reflected the correct separation of global conformational space for data points of some canonical forms where loops were close to 1 Å RMSD from the ground truth structure (Figure 5M; Supplementary Figures 6E, F).

There are only 9 data points of CDRL3 length 7, and this may explain why all predictions of CDRL3 length 8 for the starved model fell into a single cluster. In contrast, models starved of CDRL3 lengths 9 and 10 were able to separate some predictions into areas of conformational space close their ground truth structure. This may have been due the abundance of data points either side of the missing length in training data. These experiments, in addition to the structural overlap of predictions of different lengths, provided evidence of generalisation by ABB2 with origins in CDR length independence. Furthermore, these out-of-domain experiments serve as a powerful method to further explore the ability of deep learning-based structure prediction methods to extrapolate and find evidence of truly novel predictions.

Retraining whilst withholding canonical conformations highlights limited ability to extrapolate

Our analyses so far have not found any evidence of structural clusters representing novel conformations within the CDR loops of predicted antibody structures. Therefore, we set out to explicitly test whether ABB2 could predict a loop conformation not seen in the training data and without a parallel example at a different length. We ran out-of-domain experiments to train models in the absence of all examples of a specific canonical cluster and any close conformations (see methods). We focused on CDRL1 lengths 6-9 as these analyses showed the clearest cluster separation and the smallest proportion of data points that did not fall into a DBS cluster. This helped to avoid any ambiguity in the contents of the training data.

Separate models were trained for each withheld canonical class (numbers and training details given in Table 3). Each model was analysed as before, by predicting the structures of sequences in the withheld data and assessing the individual prediction accuracy as well as total occupancy of structural space.

For the 'starved' model trained in the absence of CDRL1-Len-6 canonical cluster L1-11-3 (blue dots in Figure 6A, for sequence motif see Table 1), all predictions failed to match the ground truth conformation (Figure 6A) with mean (SD) RMSD difference of 1.54 (0.51) Å. The high standard deviation reflects how some predictions were closer (less than 0.5 Å) to the ground truth structure, however the majority adopted a similar conformation to the closest canonical form L1-11-4 (pink dots in Figure 6A, cluster centroid distance of 1.81 Å) rather than the more distant forms of L1-11-1 or L1-11-2 (green and mustard dots Figure 6A, combined into one cluster by our method, centroid distance: 2.75 Å).

A more pronounced loss of the ability to extrapolate was observed for models starved of a canonical form in the remaining experiments using CDRL1-Len-7 (Figure 6B) and CDRL1-Len-9 (Figure 6D). Here all withheld data points fell into a more distant region of conformation space associated with mean prediction accuracies that were much lower, at mean (SD) RMSD values of 1.98 (0.09) Å RMSD for length 7, and 3.34 (0.19) Å for length 9.

For CDRL1-Len-7 the low prediction accuracy may have been due to very similar sequence motifs (DBS cluster: QSVSSSY, *corresponding PyIgClassify2 cluster L1-12-1: RASqSVSSSYLa*, versus DBS: QSVSSNY, *PyIgClassify2 cluster L1-12-2: RASQSVSSNYLA, see* Table 1), as well as a small number of examples within the training data (63 examples of the withheld canonical form). In the case of CDRL1-Len-8, the model made predictions of L1-13-3 (purple dots Figure 6C) that were in a similar region of structural space, however they were still 2.43 (0.19) Å away from ground truth values (Figure 6C).

These out-of-domain tests clearly demonstrated that the models trained in the absence of a canonical class were unable to recapitulate the correct conformations which could be predicted by the fully trained ABB2 ensemble. The extent of the distance between predictions and ground truth differed for each starved model and related to both sequence similarity as well as structural deviation. This led us to investigate whether prediction accuracy could be recovered by adding very small amounts of training examples back into the model.

TABLE 3 Details of retraining in the absence of canonical clusters.

	Original Model	CDRL1-6 Starved	CDRL1-7 Starved	CDRL1-8 Starved	CDRL1-9 Starved
FV structural units seen in training (including all copies in the asymmetric unit)	5771	5459	5556	5671	5554
Dropped units (corresponding to a specific canonical form)	0	312	215	100	217
Total units with CDR and length tested (before withholding a specific form)	_	2636	532	428	565
Remaining units with CDR tested (after removal)	_	2324	317	328	348

Breakdown of the number of antibody variable fragment (FV) structural units used to train ABB2 and the subsequent 'starved' models where all units containing a specific canonical form were withheld. The term structural unit is used to account for multiple copies being present in the asymmetric unit of the same PDB file. For each model, the total number of units seen in training is given, followed by the number of removed units. Then the total number of units related to the CDR being withheld is given (for example all units with CDRL1 IMGT length 6), followed by the number of those units remaining after withholding those related to a specific canonical form.

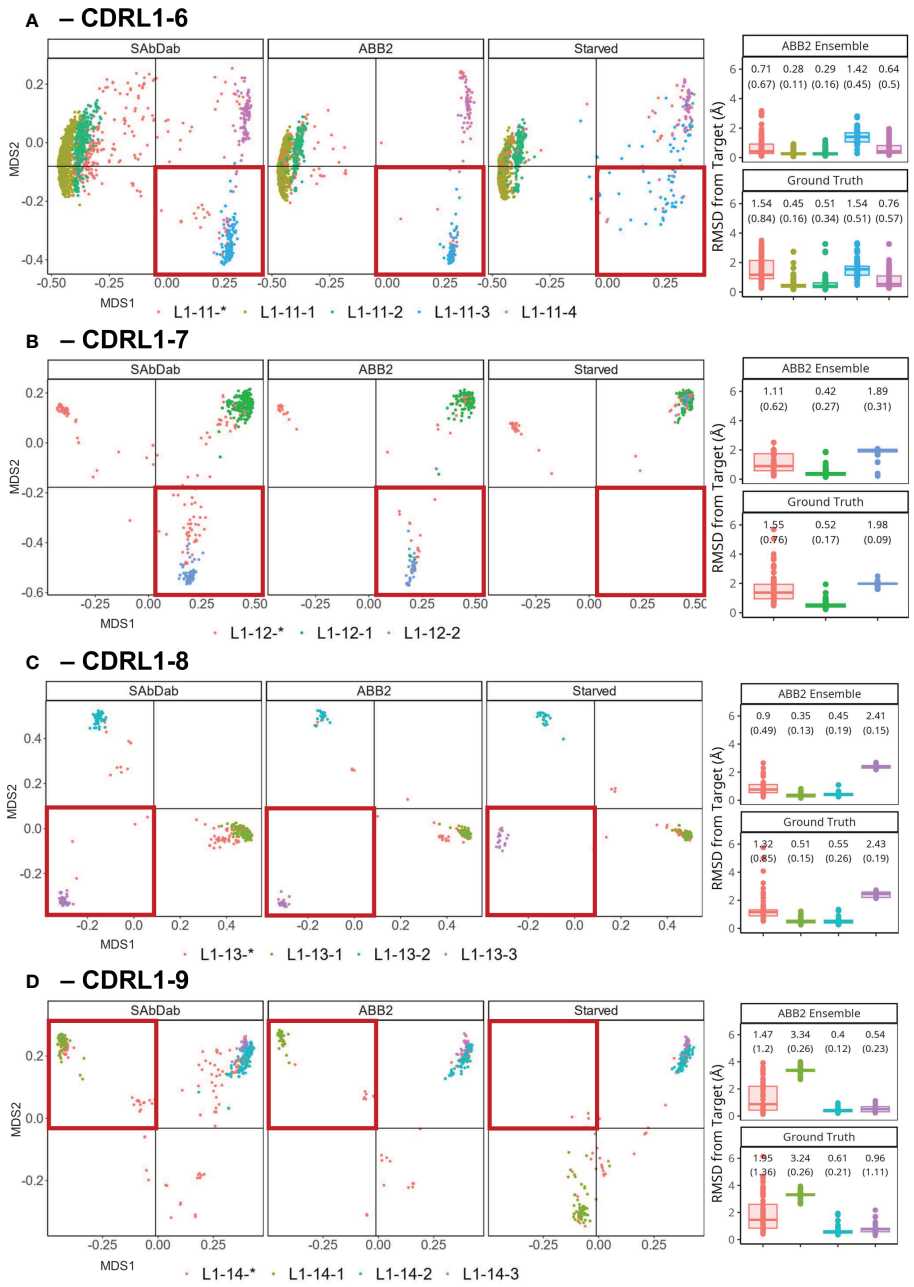

FIGURE 6
Retraining whilst withholding canonical clusters highlights limited ability to extrapolate. Results of out-of-domain experiments where ABB2 was retrained in the absence of all experimental data points assigned to a specific canonical form in PyIgClassify2 (both high and low confidence) for CDRL1 lengths 6-9 **(A–D)**. MDS plots show experimental data points coloured according to their high confidence PyIgClassify2 canonical form annotation. Any loops that do not belong to a high confidence PyIgClassify2 cluster (defined by an asterisk in the label, e.g., L1-11-*) are coloured in red. CDRL1 length 6 was retrained in the absence of all data points assigned to the PyIgClassify2 cluster 'L1-11-3' [coloured blue in **(A)**]. For CDRL1 length 7 'L1-12-2' was dropped [coloured blue in **(B)**]. For CDRL1 length 8 cluster 'L1-13-3' was dropped [coloured purple in **(C)**], and for CDRL1 length 9 cluster 'L1-14-1' was dropped [coloured green in **(D)**]. For each panel, the MDS of experimental data points found in SAbDab are shown in the far-left panel. The MDS plot of ABB2 model predictions of the corresponding sequences are shown in the middle panel (labelled 'ABB2'), and predictions of the starved model are shown in the right panel (labelled 'Starved'). The area of structural space investigated through exclusion during training is highlighted in a red box. The boxplots (median and upper and lower quartiles) and dot plots on the far right of each panel indicate the RMSD values for each predicted data point from the starved model from its target structure in predictions from the fully trained ABB2 ensemble (top graph) or the ground truth structure (bottom graph) with each sub graph labelled accordingly.

Inclusion of a small number of examples in training is sufficient to recover predictive capacity

We set out to assess whether limited amounts of training data could recover missing predictions and thus quantify the level of representation needed to produce more accurate models. We chose CDRL1-Len-6 and CDRL1-Len-7 as these had the largest standard deviations in prediction accuracy (see box plots Figure 6). For each we progressively added increasing numbers of data points back into the initial out-of-domain tests resulting in separate models for each addition of data (CDRL1-Len-6 Figures 7A, B and CDRL1-Len-7 Figures 7C, D).

pt?>As there were different total numbers of examples of CDRL1-Len-6 and CDRL1-Len-7 loops in training data (Table 3), the absolute number of PDB structures added back for each experiment did not represent the same proportion of datapoints. Therefore, we calculated the included loops as a percentage of all CDRL1 loops seen for each model (Table 4). Using these values (Table 4) and inspection of the corresponding model performance (Figure 7), we could see that very small percentages of training data (less than 1%) were enough to allow the models to accurately

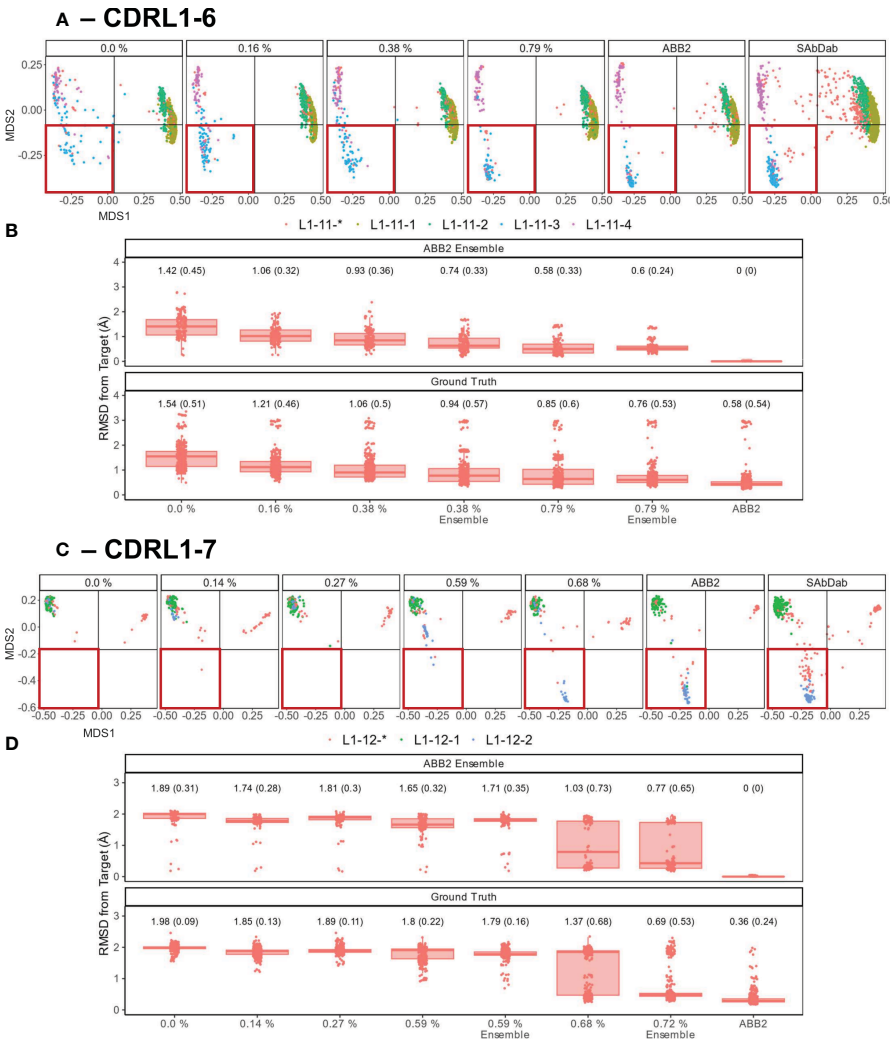

FIGURE 7
Data points that make up less than 1% of training data are sufficient to recover predictive capacity. Out-of-domain experiments were performed as in Figure 6, however for each model a specified number of experimental data points relating to the withheld canonical form were added into the training data (left to right for each panel, graphs are labelled with the percentages of data points added). Separate models were trained for each of the incrementally increasing data additions until the prediction accuracy came close to the fully trained ABB2 ensemble predictions and the ground truth experimental data. MDS plots for CDRL1 length 6 models **(A)** and CDRL1 length 7 **(C)**. The amount of data being included is labelled at the top of each panel as a percentage of the total number of unique structures out of all data points seen in training (see Table 4). The corresponding ABB2 predictions and ground truth data are shown in the far-right MDS plots in each panel. RMSD distance values are plotted in **(B, D)** for the predicted data points relating to the dropped canonical form from each model relative to the ABB2 ensemble predictions (top graph) or ground truth data (bottom graph). For each model, the mean RMSD is given above the box plots, with standard deviation given in brackets. For **(B, D)**, the top, far-right graph RMSD values are zero as the ABB2 model predictions are being compared to itself, with the comparison of ABB2 to ground truth data shown below.

10.3389/fimmu.2024.1352703

TABLE 4 Inclusion data points as a proportion of total CDRL1 data units in training.

Number of included PDB structures of withheld canonical form	CDRL1-6 Number of structural units included (percentage of all data points of CDRL1)	CDRL1-7 Number of structural units included (percentage of all data points of CDRL1)
All	2636 (45.7%)	532 (9.2%)
0	_	_
5	9 (0.16%)	8 (0.14%)
10	21 (0.38%)	15 (0.27%)
19	_	33 (0.59%)
20	43 (0.79%)	_
21	_	38 (0.68%)
22	_	40 (0.72%)

For inclusion experiments a specific number of PDB structures were gradually reintroduced to training data for a series of models. As PDB files may contain more than one structure in the asymmetric unit, the number of exact structural units is given for each experiment, as well as the corresponding percentage of all CDRL1 data points present in each training run.

recapitulate the cluster corresponding to the missing canonical form.

To check that these small proportions of data were enough to facilitate accurate predictions within the original ensemble architecture of ABB2, instead of a single model, we trained ensembles on two inclusion proportions (one above and one below the proportion where we first saw improvement ~ 0.6%, see Table 4) and analysed the resulting final output (the average structure from all four predictions). This demonstrated that the ensemble models still failed to recapitulate at the lower percentage, while the higher percentages were sufficient for the ensemble to progressively populate the missing structural space despite still being below 1% of total examples of that CDR loop (RMSD plots marked as 'ensemble' in Figures 7B, D).

These analyses underline the importance of sufficient data representation and suggest that even small numbers of datapoints can influence the predictive capacity drawn from large datasets. Ultimately the inability of models to truly extrapolate from physical principles means researchers must pay close attention to the contents of their datasets and continue to collect experimental data that explores lesser studied areas of structural space.

Discussion

In this study we analysed the predicted structures for paired sequences present in OAS generated by ABodyBuilder2 (ABB2) (20). Our data driven approach allowed identification of structural clusters within CDR loops of the same length and subsequent linkage to existing definitions of canonical forms. Our analyses aimed to explore the ability of structure predictors to enrich the experimentally defined landscape of canonical forms and identify novel conformations that would reflect generalisation or extrapolation arising from a deep learning method.

The augmentation of existing data with predicted structures enabled us to define new canonical clusters composed of heterogeneous CDR sequences which were united by the same loop backbone conformation and a sequence motif. These arose from both subdivision of areas of conformational space with uniform PyIgClassify2 annotation, as well as within highly populated areas of conformational space not assigned to any existing canonical form definition. Novel clusters of predicted loop conformations were also produced via the phenomenon of length independence (26). We observed areas of new density which had sequence enrichments identical to those of canonical forms at a shorter CDR length but contained a positionally fixed high entropy residue likely indicative of somatic hypermutation insertion. We analysed the distances between loop conformations at both the shorter and longer lengths by dynamic time warping. This revealed that different length loop clusters were indeed closer in conformational space than any of those of the same length. Out-of-domain experiments confirmed that ABB2 was able to generalise across loops of different lengths and suggested predictions were influenced by high frequency experimental datapoints seen in shorter CDR loops.

However, our analyses could not find new clusters which had no origin in training data and thus represented true extrapolation. Therefore, we performed further out-of-domain experiments by retraining our ABB2 structure predictor whilst withholding data points which belonged to specific canonical forms. These 'starved' models were then challenged with correctly predicting the unseen CDR loop conformation. We found that ABB2 retrained in this way was unable to predict conformations not seen during training, but this inability could be resolved by inclusion of a small number of examples representing between 0.5-1.0% of the total data used in development. This suggests that effective prediction accuracy by structure predictors can be achieved for conformations even when they have very poor representation in the dataset.

Our study highlights important limitations regarding the current capabilities of deep learning structure prediction tools specific to the domain of immune receptor CDR loops. Whilst numerous studies have performed out-of-domain experiments and exploratory analysis on protein folds, the conformational space of CDR loops may offer greater challenges, particularly in regions that are inherently flexible or adopt distinct structures in bound and unbound states.

These challenges emphasise that if we wish to predict outside current known structure space, new structure prediction tools that have learnt the underlying rules which govern tertiary structure, instead of just patterns in the training data, will be required. While AlphaFold2 was heralded as a huge advance in structural biology and machine learning, the higher goal of building models that can capture biophysical laws has still not been reached (46, 47). In the

absence of architectures that can extrapolate, greater amounts of training data, particularly in regions of structure space with poor coverage, may help improve predictive accuracy. Our demonstration that a small number of examples can address gaps means that a critical mass of data could be achieved to overcome the limitations of current models. However, this does place a large burden on experimental researchers to collect more data.

Finally, our results focused on an area of structural immunology relatively abundant with data and analyses, that of antibodies. As T cell receptors become more important in both immunotherapy research and the clinic, a need to better classify and understand this protein for the purpose of structure prediction may supersede that of antibodies. Therefore, the questions posed in this study may take on more relevance in a field with a relative paucity of structure and paired sequencing data (48), as well as several unanswered questions on TCR loop flexibility and comparative conformational freedom (29).

The original purpose of canonical forms was their ability to predict structure from sequence, however these use cases have been superseded by the improved performance of ML methods for structure prediction. For experimental techniques such as X-ray crystallography to be rendered redundant in immunology we must have confidence that structure prediction algorithms faithfully replicate the most important region of immune receptors, the CDR loops.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://doi.org/10.5281/zenodo.10280181.

Author contributions

AG-W: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. BA: Data curation, Investigation, Methodology, Software, Writing – review & editing. CD: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/S024093/1) AG-W is funded by Exscientia and BA was funded by Roche. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Acknowledgments

The authors wish to thank Joao Diniz Brandao Gervasio and Carlos Outeiral Rubiera from the Oxford Protein Immunoinformatics Group for their helpful comments and discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1352703/full#supplementary-material

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

2. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. (2021) 373:871–6. doi: 10.1126/science.abj8754

3. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. (2023) 379:1123–30. doi: 10.1126/science.ade2574

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res*. (2000) 28:235–42. doi: 10.1093/nar/28.1.235

5. Bordin N, Sillitoe I, Nallapareddy V, Rauer C, Lam SD, Waman VP, et al. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol*. (2023) 6:1–12. doi: 10.1038/s42003-023-04488-9

6. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. (2023) 1–4. doi: 10.1038/s41587-023-01773-0

7. Ahdritz G, Bouatta N, Kadyan S, Xia Q, Gerecke W, O'Donnell TJ, et al. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* [Preprint] (2022). doi: 10.1101/2022.11.20.517210

8. Balestriero R, Pesenti J, LeCun Y. Learning in high dimension always amounts to extrapolation. *arXiv* [Preprint] (2021). doi: 10.48550/arXiv.2110.09485

9. Fannjiang C, Listgarten J. Is novelty predictable? *arXiv* [Preprint] (2023). doi: 10.48550/arXiv.2306.00872

10. Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*. (2021) 433:167208. doi: 10.1016/j.jmb.2021.167208

11. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. (2021) 596:590–6. doi: 10.1038/s41586-021-03828-1

12. Chakravarty D, Porter LL. AlphaFold2 fails to predict protein fold switching. *Protein Sci*. (2022) 31:e4353. doi: 10.1002/pro.4353

13. Schatz DG, Ji Y. Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol*. (2011) 11:251–63. doi: 10.1038/nri2941

14. Brack C, Hirama M, Lenhard-Schuller R, Tonegawa S. A complete immunoglobulin gene is created by somatic recombination. *Cell*. (1978) 15:1–14. doi: 10.1016/0092-8674(78)90078-8

15. Alt FW, Yancopoulos GD, Blackwell TK, Wood C, Thomas E, Boss M, et al. Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J*. (1984) 3:1209–19. doi: 10.1002/j.1460-2075.1984.tb01955.x

16. Griffiths GM, Berek C, Kaartinen M, Milstein C. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature*. (1984) 312:271–5. doi: 10.1038/312271a0

17. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci*. (2014) 111:4928–33. doi: 10.1073/pnas.1323862111

18. Regep C, Georges G, Shi J, Popovic B, Deane CM. The H3 loop of antibodies shows unique structural characteristics. *Proteins Struct Funct Bioinforma*. (2017) 85:1311–8. doi: 10.1002/prot.25291

19. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* [Preprint] (2022). doi: 10.1101/2021.10.04.463034

20. Abanades B, Wong WK, Boyles F, Georges G, Bujotzek A, Deane CM. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun Biol*. (2023) 6:1–8. doi: 10.1038/s42003-023-04927-7

21. Guloglu B, Deane CM. Specific attributes of the VL domain influence both the structure and structural variability of CDR-H3 through steric effects. *Front Immunol*. (2023) 14. doi: 10.3389/fimmu.2023.1223802

22. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature*. (1989) 342:877–83. doi: 10.1038/342877a0

23. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol*. (1987) 196:901–17. doi: 10.1016/0022-2836(87)90412-8

24. North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. *J Mol Biol*. (2011) 406:228–56. doi: 10.1016/j.jmb.2010.10.030

25. Adolf-Bryfogle J, Xu Q, North B, Lehmann A, Dunbrack RL. PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res*. (2015) 43:D432–438. doi: 10.1093/nar/gku1106

26. Nowak J, Baker T, Georges G, Kelm S, Klostermann S, Shi J, et al. Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs*. (2016) 8:751–60. doi: 10.1080/19420862.2016.1158370

27. Wong WK, Georges G, Ros F, Kelm S, Lewis AP, Taddese B, et al. SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinforma Oxf Engl*. (2019) 35:1774–6. doi: 10.1093/bioinformatics/bty877

28. Kelow S, Faezov B, Xu Q, Parker M, Adolf-Bryfogle J, Dunbrack RL. A penultimate classification of canonical antibody CDR conformations. *bioRxiv* [Preprint] (2022). doi: 10.1101/2022.10.12.511988

29. Wong WK, Leem J, Deane CM. Comparative analysis of the CDR loops of antigen receptors. *Front Immunol*. (2019) 10. doi: 10.3389/fimmu.2019.02454

30. Sivasubramanian A, Sircar A, Chaudhury S, Gray JJ. Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins*. (2009) 74:497–514. doi: 10.1002/prot.22309

31. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: A resource for data mining next-generation sequencing of antibody repertoires. *J Immunol*. (2018) 201:2502–9. doi: 10.4049/jimmunol.1800708

32. Olsen TH, Boyles F, Deane CM. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci Publ Protein Soc*. (2022) 31:141–6. doi: 10.1002/pro.4205

33. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput*. (2015) 11:3696–713. doi: 10.1021/acs.jctc.5b00255

34. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PloS Comput Biol*. (2017) 13:e1005659. doi: 10.1371/journal.pcbi.1005659

35. Lefranc M-P, Lefranc G. Antibody sequence and structure analyses using IMGT®: 30 years of immunoinformatics. *Methods Mol Biol Clifton NJ*. (2023) 2552:3–59. doi: 10.1007/978-1-0716-2609-2_1

36. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. *Bioinforma Oxf Engl*. (2016) 32:298–300. doi: 10.1093/bioinformatics/btv552

37. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. (2013) 41:W34–40. doi: 10.1093/nar/gkt382

38. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. *Nucleic Acids Res*. (2014) 42:D1140–1146. doi: 10.1093/nar/gkt1043

39. Schneider C, Raybould MIJ, Deane CM. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res*. (2022) 50:D1368–72. doi: 10.1093/nar/gkab1050

40. Honegger A, Plückthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol*. (2001) 309:657–70. doi: 10.1006/jmbi.2001.4662

41. Meert W, Hendrickx K, Van Craenendonck T, Robberechts P, Blockeel H, Davis J. DTAIDistance. (2020). Available at: https://zenodo.org/records/7158824.

42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* (2011). Available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

43. Delano W. The pyMOL molecular graphics system. (2002). Available at: https://legacy.ccp4.ac.uk/newsletters/newsletter40/11_pymol.pdf.

44. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*. (2017) 33:3645–7. doi: 10.1093/bioinformatics/btx469

45. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. *J Open Source Softw*. (2019) 4:1686. doi: 10.21105/joss.01686

46. Outeiral C, Nissley DA, Deane CM. Current structure predictors are not learning the physics of protein folding. *Bioinformatics*. (2022) 38:1881–7. doi: 10.1093/bioinformatics/btab881

47. Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci*. (2024). doi: 10.1039/D3SC04185A

48. Leem J, de Oliveira SHP, Krawczyk K, Deane CM. STCRDab: the structural T-cell receptor database. *Nucleic Acids Res*. (2018) 46:D406–12. doi: 10.1093/nar/gkx971