Structural Data Recovery Through Machine Learning Integrated Gradient Analysis

Isaac G.L. Daviet

Thesis submitted for the degree of
Master of Molecular Techniques in Life Sciences

60 credits

KAROLINSKA INSTITUTET,
STOCKHOLM UNIVERSITY,
KTH ROYAL INSTITUTE OF TECHNOLOGY

May 2024

Table of Contents

Acknowledgements

This study was conducted from January to May 2024 in partnership with the Greiff Laboratory at the Department of Immunology, Rikshospitalet, Oslo and submitted as the final thesis report as part of the Molecular Techniques in Life Sciences masters program jointly hosted by Karolinska Institutet, Stockholm Universitet, and the KTH Royal Institute of Technology in Stockholm.

First and foremost, my sincerest and deepest gratitude to Dr Victor Greiff, Associate Professor at the University of Oslo and head of the Greiff Research Group, for his role as my primary research supervisor. Thank you for extending and organizing this opportunity for a young researcher to delve deeper into his passion of antibody design, and for entrusting him with the academic and educational freedom to explore the best applications of the research question at hand. Additionally, many thanks for the careful and thoughtful oversight provided on both the general research as well as the constructive inputs regarding the drafting of this report.

I would also like to extend my wholehearted gratitude to all members of the Greiff Lab and all the professional associates in the broader research group for providing the research that laid the groundwork for this study. In particular, I would like to extend a special thanks to Rahmad Akbar, Antibody Designer at Novo Nordisk, and Robert Frank, PhD candidate at the Greiff Lab, for their work in generating the datasets used, as well as their careful attention that I understood every detail in addition to the constant support and input they provided throughout the project.

A special acknowledgement to Derek M. Mason and all authors of the antibody optimization paper referenced, which provided the critical starting experimental CDRH3 sequences utilized in this analysis.

A warm thank you as well to the faculty & staff of the MTLS program, whose tireless dedication and academic excellence fostered a stellar educational that allowed for students the ability to engage in previously unexplored domains, as well as my fellow students whose tireless friendship, passion, and support inspire me to no end.

Finally, my deepest and undying love to my family, biological and chosen, and especially my mother, who’s unending love, support and passion for knowledge have been with me at every step for this opportunity of a lifetime.

Stockholm, May 2024
Isaac G.L. Daviet

Abstract

Recent advances in deep neural networks (DNN) have enabled the development of models capable of effectively identifying paratope sequences as potential binders to a given epitope. However, these models retain an inherent "black box" quality, posing challenges in understanding how amino acid sequences are classified as binders or nonbinders. Such models have demonstrated the ability to recover structural information from synthetic CDRH3 sequence libraries through a dimensionality reduction of integrated gradients feature attributes. Understanding the information recovered by these models could broaden the application of fully in-silico methods and enhance our understanding of complex biochemical interactions. However, it remains to be shown whether such models can be applied to experimentally derived sequences and recover similar amounts of structural data. Here, we demonstrate that an analysis of an experimentally derived CDRH3 sequence dataset does recover structural information similar to that of the synthetic library while suggesting that other types of information may be recovered to a greater extent than structural information. Specifically, our analysis reveals a high degree of structural overlap in isolated UMAP/PCA-based clusters derived from the integrated gradients of >34,000 CDRH3 sequences classified as paratopes to the mHER antigen. This similarity is confirmed by sequence diversity analysis, yet discrepancies in cluster RMSD point to the recovery of additional information, potentially related to residue biochemical properties. These findings highlight the potential for in vitro knowledge derivation from synthetic datasets and raise questions about the relationships between the different types of recovered information. Furthermore, they underscore the broader utility of such models beyond antigen-specific CDR sequence identification, offering insights into their biochemical and structural interactions. Overall, this study suggests that IGs analysis combined with dimensionality reduction and diversity comparison can unveil broader structural and physicochemical patterns within antibody libraries, paving the way for deeper insights into their biochemical and structural dynamics.

Abbreviations

CDRH3: Heavy-chain Complementarity Determining Region 3
DR: Dimensionality Reduction
IG: Integrated Gradients
IgG: Immunoglobulin G
ML: Machine Learning
PCA: Principal Component Analysis
PDB: PDB structural file format
UMAP: Uniform Manifold Approximation and Projection
XRAI: eXplanation with Ranked Area Integrals

Press Release: Antibodies and the Future of Synthetic Biology

Antibodies have played an increasing role in our lives - be it through the scraping of our nostrils to see if we were COVID-free, over-the-counter pregnancy tests, or even the countless new therapies promising to treat previously incurable diseases. With their lock-and-key mechanisms, the ability of these mighty weapons of the immune system to detect miniscule traces of nearly any biological substance with astonishing precision is one of the miracles of biology. Understanding them, however, is no small feat. With humans having an estimated 10^{12} antibodies unique receptors (yes, that a 1 followed by 12 zeros) studying them using traditional lab methods becomes a lengthy and expensive task. But recent breakthroughs in AI and machine learning are revolutionizing how we study and harness these microscopic medical miracles. Researchers now are developing models that can churn out millions of synthetic antibody sequences that can be used in another model to predict which ones will bind to a given antigen - just models all the way down… Yet, so far these can only predict sequences in binary terms - as binders or nonbinders to a given antigen - without revealing much about the underlying logic behind their decisions.

But if we could peek under the hood of these models? Would we be able to gain a better understanding of how these bonds work, if we could understand what these models see that we don't? It is exactly this question in mind that researchers at the University of Oslo were excited to see such that these algorithms could successfully recover biochemical information and structure - critical in understanding how an antibody binds - from both synthetic and experimentally derived antibody sequences. While we're still a long way off from fully understanding the intricate dance occurring between antibodies and antigens, these results are a crucial step in developing new computational methods that would allow us to explore these relationships with exponentially more efficiency, with the ever distant promise of fully synthetic antibody research and design slowly getting closer in the horizon.

1. Introduction

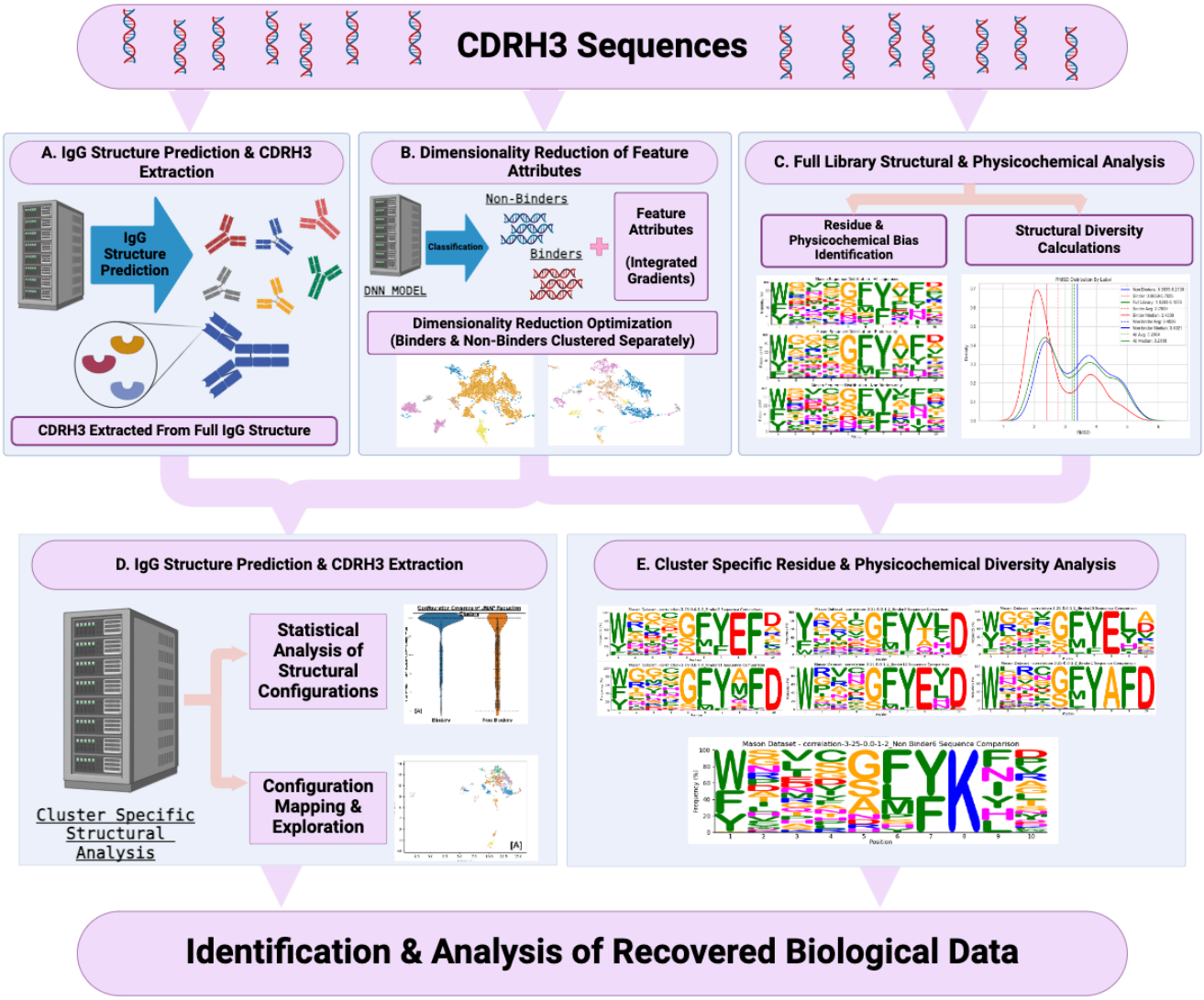
Antibody-antigen interactions are fundamental to the immune response, with an estimated 10^{12} antibodies playing a critical role in identifying and intervening against external invaders[(1)]. Within the complementarity-determining variable region, the 10 residue CDRH3 sequence is widely recognized as contributing the strongest effects on overall an antibody's binding capacity[(2–4)]. Despite significant challenges, a molecular level understanding of these interactions is essential for advancing a broad range of research[(2,5)]. With only ~7,000 of these complexes having been resolved experimentally at the atomic level[(6–8)], researchers have provided valuable but extremely limited insight into the intricate nature of these interactions[9]. Considering the vast diversity possible CDRH3 sequences, with 20^{10} possible sequences and an even greater number of potential epitope-binding sites[(10)], this stark scarcity of structural and affinity data[(11,12)] poses significant research bottlenecks and underscores the necessity for novel methodologies that enable the generation of large datasets to enable detailed molecular analysis of these paratope-epitope interactions[(12)].

Recent advancements in ML have significantly enhanced the potential of in-silico methods to address these problems[13]. Although primarily used as rough binary binding prediction tools[12–14], their ability to decode the hidden nonlinear rules governing antibody-antigen interactions[15–21] suggest that an in-depth analysis of the underlying calculations could play a role in elucidating these interactions[22–24]. Despite this potential, the refinement of such models has suffered from similar limitations due to the small and incomplete datasets typically available, often comprising fewer than 10,000 antibodies[9].

Addressing this data scarcity, the Absolut! suite has been developed to allow for the creation of datasets containing millions of CDRH3 sequences through lattice based 3D antibody-antigen structures, which can be used to train and refine deep neural network (DNN) models to categorize binder/non-binder sequences with remarkable efficiency[25]. Moreover, the analysis of the model's integrated gradients (IG) - values which quantify the importance of each input feature in the decision making process[26,27] - has been demonstrated to recover some structural data from raw sequences[25]. Validation of these results would unlock new possibilities for understanding the complexities of protein-protein interactions while demonstrating that the input of explicit structural information into such models may not be necessary, potentially leading to significant improvements in speed and computational efficiency.

Building upon these results, this study aims to apply and validate methodologies on experimentally derived data, having as key objective to assess the ability of DNNs to recover structural and other biological data from such datasets through dimensionality reduction (DR) analysis of integrated gradients, and thus derive methodological recommendations for future research. In particular, we sought to demonstrate that such models could transcend raw sequence input in their classification strategies by confirming high sequence diversity within reduction clusters.

As a build up towards the work of this study, an experimentally derived anti-HER2 CDRH3 library described by Mason et al.[28] was classified as binders or nonbinders using one such model, with IG values calculated for each sequence[25]. Subsequently, these sequences were utilized to predict CDRH3 structures through IGFold[29]. Preliminary analyses included a comprehensive review of sequence diversity and RMSD ranges across the entire library, followed by the optimization of UMAP and PCA reductions, testing a variety of components, parameters and distance metrics. Significant clusters were further processed using SPACE2[29] to identify structural configurations. These were then matched to the appropriate sequences to allow for the visualization of structural patterns on the initial reduction graphs. The study culminated in an examination of sequence and physicochemical diversity within these clusters to determine if any further recovery patterns could be identified.



Graphical Abstract | *This study, utilizing an experimentally derived CDRH3 sequence library composed of high & low affinity HER2 binders, analyzed the recovery of structural and physicochemical data through (C) a full library analysis to identify any residue/physicochemical biases and estimate RMSD range. The sequences were then (A) used to predicts full IgG antibody structures using IGFold, with the CDRH3-specific structures then isolated for further analysis. (B) Concurrently, the sequences were passed through the DNN binary classification model developed in Robert et al.(25) which categorizes the sequences as binders or nonbinders to the inputted HER2 antigen, from which the integrated gradients are extracted and used in a dimensionality reduction optimization to extract separate binder/non-binder clusters. (D) Structural configurations for each unique cluster are then identified, analyzed and mapped onto original reduction graphs. (C) Concurrently, a full library analysis is conducted to identify any residue or physicochemical biases present, and to determine the overall RMSD range present in the library, (E) which are then compared to a similar cluster-specific analysis using the same clusters derived from (B). All results are then compared to identify and analyze potential structural and physicochemical data recovery from the DNN model. Created in BioRender*

2. Methods

2.1. Prior Work & Data Generation

Previous work by Mason et al. provided the foundational CDRH3 sequences through deep-sequencing an anti-HER2 library[28]. Prior to this study, these binder sequences were converted into one-hot-encoded format by the Greiff lab and labeled as binders or non-binders using the DNN model described by Robert et al[25]. Following this, IG feature attributes values were extracted for each input sequence, resulting in each residue space having an value describing the overall importance of that residue type at that space for that specific sequence in the model's final classification of said sequence. Given that all sequences analyzed are high or low affinity HER2 binders and the exploratory nature of the study, the specific classification efficacy of the model was not considered past what was described in this article. Rather, this exploratory study was concerned with analysis and visualization of potential biological data recovery patterns arising from classification.

Additionally, the raw sequences were utilized to construct full IgG models, varying only the unique CDRH3 regions using the structural prediction tool IGfold[29]. This comprehensive data generation resulted in three distinct datasets comprising 34,146 sequences - 8,955 labeled binders and 25,191 as nonbinders - comprised of the labeled raw sequences, one-hot-encoded IG data, and full IgG PDB files for this study's analyses.

2.2. Preliminary Dataset Analysis

2.2.1. Library Wide RMSD Calculations

In the preliminary analysis of the dataset, RMSD values were calculated for randomized subsets drawn from the complete dataset, including subsets composed of only binders or nonbinders to ascertain the structural variances within the CDRH3 library. This separation allowed for an early quantification of structural biases present in the library, critical to understanding the more focused structural analysis conducted in later sections.

To achieve this, randomized subsets of 8955 CDRH3 sequences - the total number of binders -were generated from the full library, binders-only and nonbinders-only subsets for calculation consistency. From this, two sequences were randomly selected for RMSD calculation and removed from the dataset. Utilizing the aggregated RMSD values, the minimum, maximum, median and mean RMSD were derived for each subset, offering a preliminary understanding of the structural variance present in each subset type. Given the nature of these calculations, this process was meticulously repeated thrice for each of the data subsets (full library, binders-only, and nonbinders-only), confirming result consistency while maximizing computational efficiency. It should be noted that the results obtained from such a method will contain a relatively high degree of variance between resulting ranges dependent on which of the sequences were sampled, however the relative means and medians remained fairly consistent. This relatively simple set of calculations allowed us to visualize and calculate the ranges and summary statistics for each subset (Figure 3) as a point of comparison to the cluster specific SPACE2 structural analyses performed in later sections.

2.2.2. Library Wide Sequence Diversity Assessment

To further investigate the sequence diversity present within the library, we generated logo plots for each of the previously described subsets. These visual representations allowed us to scrutinize the variation in sequences and pinpoint biases inherent to the dataset through identification of high or low residue variability positions. This step is crucial in delineating regions of potential interest within each subset. Furthermore, the distribution of residue physicochemical properties was also examined across all subsets through coloring based on the physicochemical classifications described by Lesk[30] (Figure 4), which will prove crucial in further inferring inherent library biases and recognizing regions or patterns of interest within later analysis.

2.3. Optimization of Dimensionality Reduction Techniques

2.3.1. Selection and Adjustment of PCA and UMAP Parameters

To refine visualization of high-dimensional IG data, various PCA and UMAP DRs configurations were evaluated to generate clusters with distinct distribution patterns between classification labels (Figure 1A/C/D). While generally considered less reliable and prone to bias due to the heavy optimization required, exploration of UMAP reductions was largely informed by their ability to consistently generate small clusters with significant segregation between the binder and non-binder distributions. This would allow us to easily calculate a large number of these clusters for more focused analyses. As will be discussed in later sections, the PCA reductions did not display such point distributions, requiring the analysis of much larger clusters. However, given the overall simplicity and resulting reliability of this technique, it was decided to include some PCA graphs in the analysis so as to ascertain if some recovery patterns could still be identified.

Beginning with UMAP, no prior research could be found regarding the optimal distance metric for this type of IG data. Therefore, an exploration of the cosine, correlation, Euclidean, Manhattan, Hamming and Jaccard distance calculation metrics was conducted to ascertain which is most suitable. For each of these, the minimum distance and number of neighbors parameters were optimized based on point distribution of the first 2 components to generate a selection of parameters that exhibited significant separation between clusters as well as segregation between labels. Using these selected parameters, a final examination was conducted on the clustering patterns exhibited by the first 5 components to isolate for further examination the UMAP reductions that exhibited the strongest label-segregated clustering.

Concurrently, a separate PCA reduction optimization was conducted. Given its relative simplicity and the contrast in point distribution compared to that seen in UMAP, notably the distinct spike-like distribution making cluster identification difficult (Figure 1B), efforts were primarily focused on identifying principal components with distinct variance between the labels. To that end, the number of components required to account for 99.99% of the total explained variance for both labels were determined. A pairplot of all the components up to the threshold was then generated and analyzed to isolate a select few components with asymmetrical distribution patterns, which will be used for further analysis.

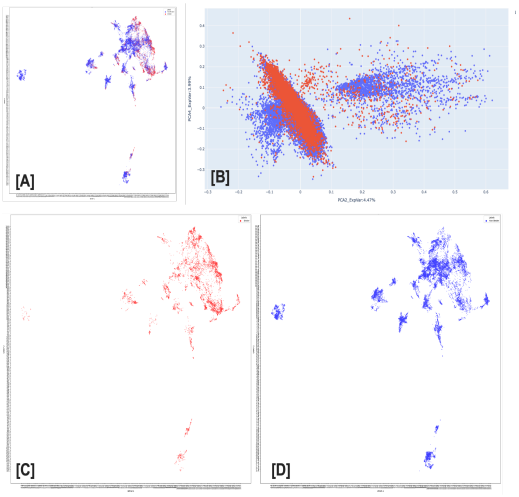


Figure 1. PCA reductions exhibited minimal segregation between binders/non-binder sequences compared to optimized UMAP reductions. *(A, C, D): Example of optimal final UMAP reduction use for analysis, where clear and significant differences can be identified between binders (red) and nonbinders (red). While still following similar patterns, significant differences in binder (B) vs non-binder (C) densities can still be identified [A], allowing for simplified cluster identification and extraction. Parameters: metric = correlation; n_components = 3, n_neighbors = 25; minimum_distance = 0.0; components 1 & 2. (B): Example of final PCA reduction use for analysis, demonstrating the overall lack of segregation between labels characteristic of PCA, with binders in particular exhibiting minimal distribution range compared to nonbinders, making cluster identification and extract difficult. Principal Components of graph are 2 & 4*

2.3.2. Evaluation and Cluster Extraction of Dimensionality Reductions

Upon selection of the various reductions, the clusters of interest were extracted separately to simplify our analysis, resulting in a plethora of separate binder and non-binder clusters. The UMAP point distributions allowed for the automated calculation of the clusters using the DBScan clustering algorithm[31] (Figure 2), following a separate optimization of the epsilon and minimum samples parameters required for each reduction. This resulted in all the sequences of the smaller, more isolated clusters to be extracted while simultaneously separating denser point distributions into smaller distinct clusters.

Once extracted, each UMAP cluster was assigned a priority level based on the percentages of sequences it contained relative to the total sequences within the label-specific dataset, with ‘low’ indicating a cluster containing less than 1% of the total subset, medium indicating 1-10%, ‘high’ indicating 10-40%, ‘out-of-bounds’ indicating >40%, and ‘unclustered’ indicating sequences that remained unclustered by the DBScan algorithm. It should be noted that these thresholds were set somewhat arbitrarily and serve mostly as a method of data organization and for ease of computation. For example, this study excluded all ‘unclustered’ sequences from the structural analysis so as to avoid wasting processing power on outlier points with an extremely broad distribution across entire reductions, which would have caused difficulties in later analyses.

On the other hand, the challenging spike-like point distributions patterns of the PCA reductions, with points emanating in a small number of densely packed ‘spikes’ emanating from a relatively central point (Figure 8B), were clustered through a visual approximation of a broad set of vertices which grouped sequences according to the most visually distinct distribution patterns in each reduction for a manual extraction. This typically yielded one to two large clusters for binders and two to four for non-binders per PCA reduction.

It should be noted, however, that the main purpose of these clusters was to separate the sequences into more manageable arbitrarily defined groupings. As discussed in later sections, this step may not be

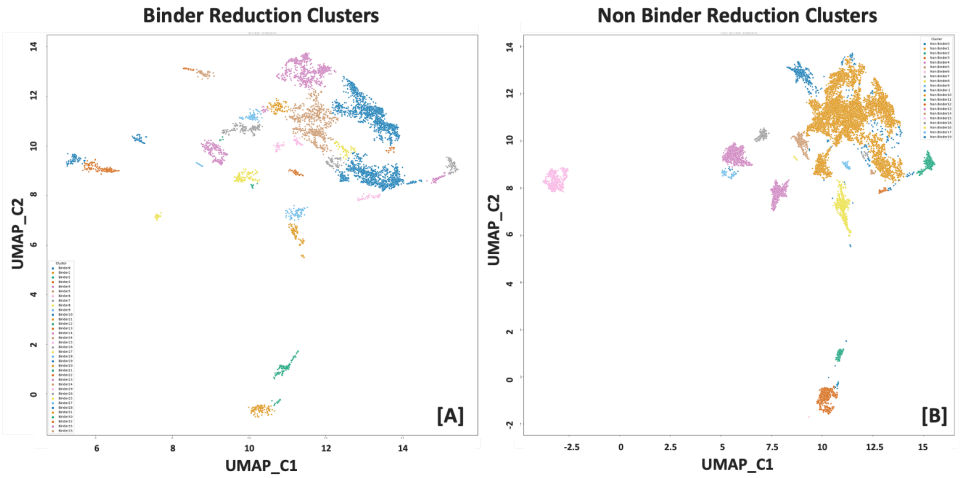


Figure 2. DBScan clustering of UMAP reductions allowed for efficient extraction and evaluation of CDRH3 sequences, particularly for binder sequences. *DBScan clustering results of example reduction from Fig.X, displaying clear and significant differences clusters for both binders (A) and nonbinders (B). Clusters extracted using the same parameters (epsilon = 0.15; min_samples = 20) for both label types, explaining the differences in the number of clusters identified at 33 for binders and 19 for nonbinders due to differences in point densities.*

necessary for future PCA analyses, with a structural comparison of the full dataset probably being more appropriate. However, given our interest in focused analysis of smaller clusters and our desire to compare UMAP to PCA, this step was conducted in this project for the sake of consistency.

2.4. Structural Cluster Identification and Analysis

2.4.1. Implementation of SPACE2 for Structural Clustering

In-depth structural analysis was conducted using the SPACE2 clustering algorithm[29] to discern patterns within the isolated reduction clusters and identify structural similarities, focusing exclusively on the predicted CDRH3 regions rather than the full synthetic IgG structures so as to prevent structural clustering on the basis of the other sequentially identical regions. Additionally, as mentioned in the previous section, reduction clusters were extracted so as to contain only binders or nonbinders. Therefore the resulting structural analysis considered only binders or nonbinders from the same cluster, rather than performing a structural comparison of the entire dataset against itself so as to explore the structural diversity within individual reduction clusters.

Given the computational demands of SPACE2, and in order to maximize the number of CDRH3s in epitope-consistent multiple-occupancy clusters, we adhered to the recommended default settings for all uses of SPACE2 in this study, which calculates structural clusters using an agglomerative algorithm, with an RMSD cutoff of 1.25 and the number of jobs set to 1[32]. These result were systematically compiled into structured dataframes, allowing an in-depth comparative analysis both within and across different reduction clusters

2.4.2. Analysis of Structural Cluster Distribution of RMSD values within Reduction Clusters

Once compiled, the above mentioned data frames were used to analyze the diversity of structural configurations in a given cluster. This included the calculation and visualization of the RMSD values of the multiconfiguration reduction clusters, allowing us to evaluate the degree of structural similarity and ranges among clusters containing multiple structures (Figure 5, 11). Additionally, the percentage each structural configuration covered of its associated reduction cluster was also calculated and visualized to further gain an understanding on the overall structural diversity of clusters, as well as quantify the effectiveness and accuracy of the different reduction methods used (Figure 14). While such analyses will not provide conclusive evidence towards proving or disproving the aims of this study, important contextual information can still be retrieved and used to further support or disprove the overall observations and conclusions.

2.5. Overarching Structure Identification and Analysis

2.5.1. Consolidation of Structural Data Across Reductions

To derive a comprehensive view of structural diversity within the dataset, the structural configurations calculated for each sequence were amalgamated by distance metric (for UMAP reductions) or component (for PCA reductions), as well as a subset containing all structural configurations across all factors. Duplicate configurations were then removed, and each subset was submitted for another round of SPACE2 clustering. This allowed for the confirmation that the observed configurations identified by each metric or component for each sequence were consistent between and across factors. Additionally, this provided us with a single consistent configuration name for each sequence, enabling the configuration mapping discussed in later sections. This consolidation was necessary due to the processing of the SPACE2, which names a given configuration grouping based on the first sequence inputted. Thi step was required as, depending on the reduction, the makeup and order of a cluster could differ slightly and result in a different name for what would ultimately be considered the same configuration grouping.

2.5.2. Comparative Analysis of Structural Configuration Distribution Across Reductions

Following the configuration consolidation described above, a comparative analysis was conducted to assess the consistency and variability of these final configurations across all selected reductions, distance metrics and components. This analysis refined the overall structural model of the library by determining the stability and relevance of the superclusters, followed by an exploration of the overlaps between the final structural configurations and original reduction clusters to understand if and how these characteristics are correlated. This combined approach would provide conclusive evidence towards the model's ability to recover structural data, while also providing further contextual information critical to the final evaluation of the DR techniques used.

To that end, the metric-specific as well as the overarching final configurations, were mapped onto all of the original reduction graphs. This included a mapping of the UMAP derived superclusters onto the PCA

reductions and vice-versa so as to compare the overall distribution of structural configurations. Given the differences in the point distributions and the resulting difficulties in full dataset analyses of the PCA graphs, this allowed us to confirm the structural data recovery of each technique (Figures 6B/D & 7B/D) Additionally, separate mappings were produced that distinguished whether each sequence supercluster was identified by at least one metric/component or all tested metrics/components. This systematic approach allowed for the visualization of the consistency of structural configurations across different analytical dimensions and to determine the retention rate of structural information across all reduction methodologies, facilitating a nuanced evaluation of the recovery of structural data across all tested DR techniques.

2.6. Detailed Analysis of Sequence & Physicochemical Property Diversity Across all Clusters

2.7. Analysis of Residue Distribution Variability

Upon completion of the structural mapping described above, a detailed examination of residue distribution across all structural and reduction clusters was conducted to identify patterns of variability and conservation within the CDRH3 sequences (Figure 10). This analysis highlighted specific positions characterized by high or low residue variability, which were then compared to the results of the library diversity analysis (Figure 4). Furthermore, the distribution variability within binder and non-binder clusters was also compared to explore how sequence diversity may influence binder classification (Figures 10-12). This multi-pronged comparative approach provided insights into the implications of sequence variability of CDRH3 sequences by cluster, with the primary goal being the confirmation that DNN model classification does not occur not solely based on sequence similarity.

2.8. Analysis of Physicochemical Properties Biases

In conjunction with the residue distribution analysis described in the previous section, a similar analysis was conducted on the distribution of physicochemical properties across all structural and reduction clusters and compared to the full library analysis. To that end, we focused primarily on the categorization of hydrophobicity, charge and size, as delineated by the Lesk coloring method. While somewhat outside the scope of this study, the goal was to potentially identify intriguing patterns in their distribution, which may indicate a degree of non-structural data recovery. This was conducted with future research in mind by providing preliminary insights into if physicochemical properties could also be recovered, and how such information interacts with any recovered structural data.

2.9. Comparative Evaluation Of Dimensionality Reductions

A post-analysis comparison of the different DRs techniques was conducted to identify what broader conclusions could be derived to identify which combination of method, metrics, components and parameters best captured the underlying structural data, with particular emphasis being placed on their ability to generate meaningful overlap of reduction clusters with structural configurations and sufficient sequence capture efficiency while maximizing computational efficiencies. These findings were then documented to guide methodological choices in future research endeavors.

3. Results

3.1. Comprehensive Library Analysis

3.1.1. Higher Observed Overall Structural Similarity Between Binders

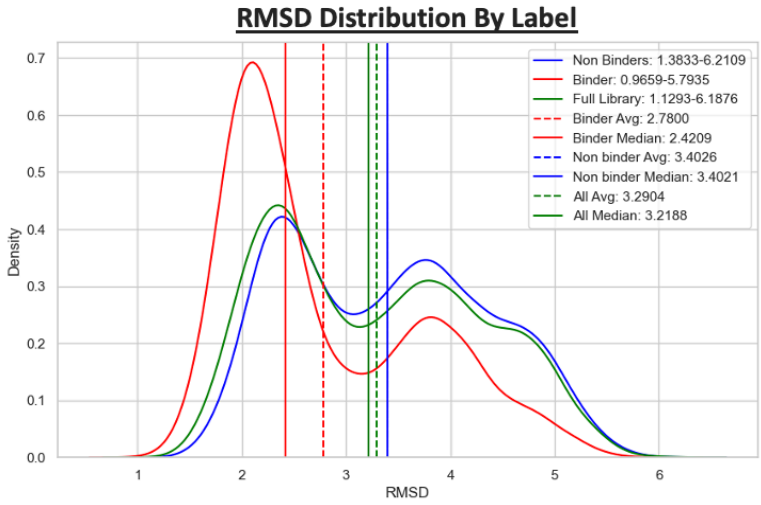


Figure 3. ***Binder-classified sequences exhibit an overall greater degree of structural similarity than both nonbinders and the full dataset, despite broad similarities in range.*** *RMSD calculations of the three library datasets (full, binders-only, nonbinders-only). Full library (green) range: 1.29-6.19; mean: 3.29; median: 3.22. Binders (red) range: 0.97-5.79; mean: 2.78; median: 2.4. nonbinders (blue) range: 1.38-6.21; mean: 3.40; median: 3.40*

Our analysis of the RMSD averages across the different subsets of the full mason HER2 binders library analysis revealed interesting and distinct patterns. Specifically, the binders-only datasets exhibited lower mean and median values of 2.78 and 2.42 respectively, compared to the full library dataset mean/median of 3.29 and 3.2, and a consistent mean/median of 3.4 for nonbinders, implying a certain degree of structural similarity across binder sequences (Figure 3).

This can be further corroborated by the notable differences in the minimum RMSD values calculated, with the binders-only dataset recording the lowest minimum RMSD at 0.97, significantly lower than the nonbinders-only datasets at 1.38. Interestingly, the binder subset calculated a maximum RMSD of 5.79 indicating that substantial structural heterogeneity is still present among binder sequences. These results, and particularly the RMSD maxima and minima, should be approached with caution due to the methodological constraints of the RMSD calculations. Nonetheless, these consistent results are the first indication that the model may be recovering some structural information from the raw sequences.

3.1.2. *Consistent Sequence Biases Identified Across All Data Subsets*

Our systematic analysis of sequence diversity revealed consistent biases across the entire library, irrespective of label (Figure 4). Residue diversity is notably lower at positions 1, 6, and 7 across all datasets, highlighting a uniform pattern of limited variability at these specific sites. Conversely, positions 2-4 and 8-10 displayed heightened diversity across all datasets, with positions 2-4 exhibiting particularly high variability. Position 5 presented a distinct case, exhibiting the most notable difference between the subsets with a clear dominance of glycine in the binders-only subset, compared to the relatively high variability in the other two.

Overall, while some differences in residue distribution can be seen at these high-variability positions, no broader label-specific trends can be derived from these results. However, the identification of these biases within both the full library and the subsets provides crucial context for subsequent cluster analyses by underscoring the relative uniformity in diversity across the subsets, essential for the interpretation of later results.

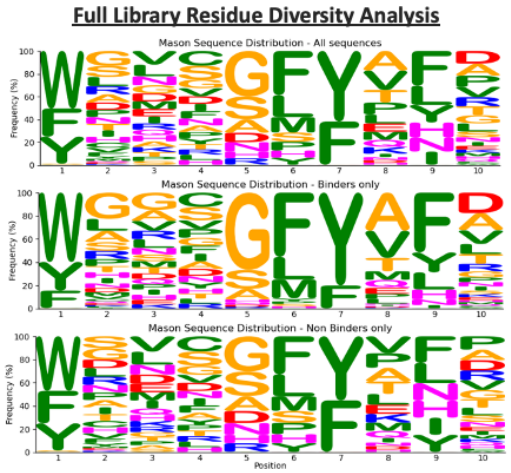


Figure 4. ***Library exhibits similar residue and physicochemical bias across all subsets.*** *Top: Full library residue diversity (34,146 sequences); Center: Binder residue diversity (8,955 sequences); Bottom: Non-binder residue diversity (25,191 sequences). Clear reduced variability at positions 1, 6 and 7, with positions 1 and 7 entirely dominated by hydrophobic residues while charged residues are almost entirely present within the high diversity positions 2-5 and 8-9 across all positions and data subsets. Apart from a slight increase in small non-polar residues (particularly glycine) is seen at binder position 5, virtually no significant differences can be observed with full dataset diversity*

Lesk color scheme[30]: Orange: Small nonpolar (G, A, S, T); Green: Hydrophobic (C, V, I, L, P, F, Y, M, W); Magenta: Polar (N, Q, H); Red: Negatively charged (D, E); Blue: Positively charged (K, R)

3.1.3. *Overlaps Between Residue/Physicochemical Distribution & Reduction Clusters*

Analysis of all plots generated using the Lesk-color scheme[30] (see Figure 4 for definition) provided identifiable physicochemical distribution patterns, allowing for the visualization of significant overlaps between the residue and physicochemical diversity. Given the limited nature of this exploration, focusing on this scheme allowed for a much easier comparison of the full dataset biases to the isolated reduction clusters. By focusing on this approach, notable physicochemical biases were identified within the library, which revealed a predominance of specific amino acid characteristics at the position exhibiting reduced diversity described in the previous section (Figure 4). These positions predominantly featured strong biases of residues classified as hydrophobic or small non-polar. It should additionally be noted that the comparatively diverse positions 8 and 9 are also dominated by these residue types, particularly in binders.

This strong bias can be clearly seen across all subsets of data as well as the full dataset. This points to these residue and physicochemical patterns being one or more motifs which are important in the binding of the sequences in the anti-HER2 library. It is also noteworthy that position 5, which was the only position exhibiting a strong sequence bias unique to the binders-only dataset, exhibits an even strong physicochemical bias in binders, being almost entirely dominated by small non-polar residues, while the nonbinders exhibit a much stronger presence of charged and hydrophobic residues.

As with the results of the sequence diversity analysis, the patterns described provide an early indication that some kind of biologically relevant data is being recovered by the model, with classification not occurring exclusively based on sequence similarity. In particular, the observation at position 5 points to the possibility of physicochemical information being recovered, which would raise further questions to the relation between this and any confirmed recovery of structural data.

3.2. Detailed Analysis of Reduction Clusters

Note: the results derived from the PCA reductions did not provide substantial enough evidence to make any significant observations or conclusions due to our inability to adequately isolate small enough clusters (Figure 13 for more details). While a more in depth analysis of the PCA results and differences between the two reduction methods will be conducted in the later sections, this section will be looking exclusively at the UMAP clustering data.

3.2.1. Reduction Clusters Exhibit Significant Structural Homogeneity

The preliminary analysis of UMAP reduction clusters manifested marked structural consistencies within both binder and non-binder clusters (Figure 5). Notably, binder clusters tended to be dominated by a singular structural configuration, suggesting a strong correlation between common structural motifs and the model's classification of a sequence as a binder. On average, each calculated structural cluster covered 88.91% of a binder cluster, with a median of 100%, demonstrating high structural consistency within each binder cluster (Figure 5A). Non-binder clusters displayed a similar but lesser pattern, with an average coverage of 70.96% and a median coverage of 92.45% (Figure 5A). These results provide the first concrete evidence that such IG derived clusters are strongly correlated with structural similarities between sequences, demonstrating the model's ability to recover structural data to an extent.

3.2.2. Multi-Configuration Clusters Exhibit Differences in Configuration Distribution

Despite the positive results described in the previous section, attention should be given to the inconsistencies in structural coherence found within reduction clusters. Specifically, a significant proportion of clusters also exhibit a high degree of internal variability and must be accounted for. This observed high variability points to a complex interplay between CDRH3 structural attributes and IG clustering. Some of these multiconfiguration clusters, for example, contained a dominant structural

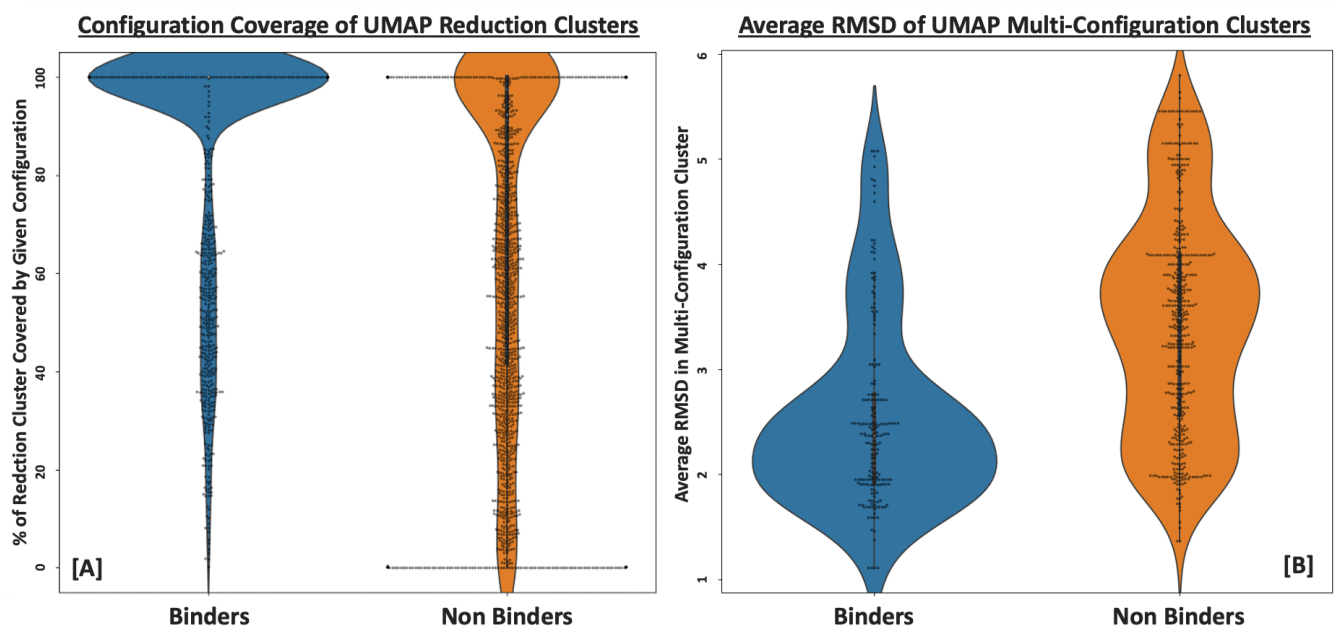


Figure 5. UMAP binder reduction clusters exhibit greater structural similarities than non-binder clusters. *(A): The percent amount of an associated reduction cluster a given configuration represents, for all reduction clusters across all UMAP reductions. 1740 Binder configurations (blue) analyzed with mean of 88.91 ± 22.96% & median of 100%, demonstrating that the majority of binder clusters contain a single configuration. Compared to 1904 non-binder configurations with mean of 70.96 ± 34.63% & median of 92.45%, indicating that while still exhibiting some degree of structural uniformity, the majority of clusters contain more than one structural configuration (B): The average RMSD value of the CDRH3 contained in all multi-configurational clusters across all UMAP reductions. 192 Binder multi-configurational clusters analyzed with mean of 2.54 ± 0.89 & median of 2.37 compared to 447 non-binder multi-configurational clusters analyzed with mean of 3.44 ± 1.04 & median of 3.49. Results and visual analysis of resulting violin plot distributions demonstrate that even binder clusters containing multiple configurations share more structural similarities than non-binder clusters. See (see Supplemental Data: appendix_folder/data/UMAP/mason_umap_SPACE2_results.xlsx for raw data*

configuration covering >80% of the cluster and a minor configuration covering <20%, while others demonstrated a relatively equal split between 2 or more clusters (Supplemental Table 2 & 3). These inconsistencies suggest that sequences are not being clustered on structural properties alone, although no conclusions can be made until the cluster mapping analysis. For example, these mixed clusters could be caused by the extracted reduction clusters being too broad, with the individual configurations having clearly distinct distributions within said clusters.

3.2.3. Multiconfiguration Exhibit Broad RMSD Ranges

Given the observations made in the previous section regarding significant structural inconsistencies within reduction clusters, special attention needs to be placed upon the RMSD analysis of these different structural configurations. To that end, it is notable that the RMSD mean/median of multi-structural binder clusters were calculated at 2.54 and 2.37 respectively (Figure 5B). These distributions more closely approximated RMSD values of the binders-only datasets than the SPACE2 RMSD cutoff, hinting at an element of structural diversity. A distribution analysis of RMSD values present within reduction clusters reveals a surprising degree of heterogeneity for both binder and especially non-binder clusters, further supporting the implications made in the previous sections.

3.2.4. Greater Degree of Inconsistencies in Non-Binder Clusters

Further complicating the results and observations made in this section are the stark differences that are consistently seen among non-binder clusters. As noted in all previous sections, non-binder clusters exhibit consistently lower indications of structural coherence (Figure 5). The median coverage value of 92.45% observed is particularly noteworthy as it indicates that the majority of the identified non-binder clusters are structurally diverse, compared to the majority of binder clusters that are structurally consistent. This could potentially have been explained by the significantly larger number of non-binder sequences resulting in substantially more structural configurations representing a more accurate distribution. Indeed, a total of 1904 non-binder configurations were identified compared to 1740 for the binders. While a substantial difference, this does not seem to be a significant enough increase to explain the differences in coverage distribution between the two classes, with the nonbinders exhibiting a much broader overall range as seen in Figure 5.

Additionally, these observations extend to the even starker differences observed between the internal RMSD value distributions. As mentioned in the previous sections, multi-configuration binder clusters demonstrate an average RMSD of 2.54 (Figure 5B), with a distribution a gradual tapering of RMSD values higher than that average. When compared to this, non-binder clusters displayed a higher average of 3.49 with a significantly broader range and point distribution, implying a differential treatment by the model or clustering methodology.

The possible interpretation that these differences could be explained by clustering differences inherent to the different UMAP distance metrics tested is undercut by the observation of similar patterns across all metric-specific calculations (Figure 11), indicating that this is indeed a broader trend in the overall anti-HER2 library. To address this, special attention will therefore be given in the following sections regarding the distribution of the structural configurations within binder versus non-binder clusters. Should these differences be further substantiated, it would provide significant evidence towards the model recovering differential amounts of structural information based on a sequence's ultimate classification.

3.3. Identification of Unique Configurations for Each Sequence

Note: While the results discussed in this section broadly apply to all reductions examined, for the sake of consistency and brevity, all figures and tables referenced in this and section 3.4 were derived from a single graph for each DR method that best demonstrates these observed results. The PCA figures were derived through analysis of principal components 2 and 4, while those of UMAP were derived using correlation distance with the following parameters: correlation; n_components = 3; n_neighbors = 25; minimum_distance = 0.0; components 1 & 2. Results from the remaining reductions can be found in the supplemental materials (see appendices for more information)

3.3.1. Structural Consistency Across Sequences Confirms Methodological Approach

Both the UMAP and PCA data were successfully able to derive a single structural configuration for each sequence, demonstrating a convergence across reductions (see Supplemental Data: *appendix_folder/data/UMAP/mason_umap-space2_superclusters.csv & mason_pca_space2_superclusters.csv*). This confirms the broader practicality of our methodological approach in determining broad sequence structural configurations derived from an in depth exploration of different reduction techniques and parameters. In total, only two sequences were determined to have a multi-structural supercluster using UMAP data, containing two structural configurations instead of one, while no such superclusters were identified using PCA data.

Interestingly, this was the main area where we observed consistent differences between the different UMAP metrics tested, although the differences were not stark. For the vast majority of sequences, all metrics were capable of identifying the same final structural configuration for each sequence (see Supplemental Data: *appendix_folder/data/UMAP/mason_umap-space2_superclusters.csv*). However, some metrics failed to identify overall superclusters for a few sequences, most likely because these sequences were unclustered in all resulting reductions for those specific metrics. While a more detailed comparison between metrics will be conducted in a later section, the highest library coverage was seen using correlation at 34,138 sequences identified and Manhattan being the lowest at 27,384 (Table 1). From a methodological standpoint, this not only indicates the importance in the proper selection of distance metrics, but also the possibility of using multiple metrics as a redundancy to capture outlier sequences that may not be properly captured even by the most efficient distance metric due to subjective judgements during cluster extraction.

Similarly, superclusters derived from PCA data were identified for all sequences with no exceptions. The main difference identified between these and the UMAP results was the complete agreement across all components in supercluster identification. In contrast to UMAP where some sequences would be significant outliers under specific distance metrics, a sequence that would be classified as an outlier under one PCA component would be classified as an outlier across all components. It is important to note, however, that these values are results of decisions made regarding decision in the extraction and analysis of the clusters, as well as the order in which each CDRH3 was processed by the SPACE2 algorithm, and are thus not indicative of structural patterns as a whole. Rather, these results comment on the efficiency of each methodological choice at identifying structural configurations across the full library.

3.3.2. Significant Overlap of Configurations With Reduction Clusters

Upon mapping the superclusters back onto the original reduction graphs, a discernable segregation based on structure became readily apparent (Figure 6-9), demonstrating a high degree of overlap between the structural and UMAP reduction clusters and confirming the recovery of structural data through the DNN IG analysis. This broadly confirms the model's ability to recover to a certain extent the structural data from an experimental library.

In so doing, incongruities consistent with results described in previous sections can also be observed. As hypothesized, some of these can indeed be explained by suboptimal cluster calculation and extraction of the initial reduction clusters, with a discernable demarcation between the structural configurations (Figure 9). However, some these multi-structural reduction clusters do not exhibit such a clear delineation between structural configurations, rather displaying a balanced distribution mix between the associated structural clusters. This was observed more frequently in the larger binder clusters, usually containing an even blend of two configurations. For non-binder clusters, this effect is particularly pronounced, with the large central clusters typically exhibiting a far more chaotic distribution of structural configurations with only a handful of large clusters further from the core distribution displaying a high degree of structural consistency (Figure 6A).

When considering the higher-than-expected RMSD values of these structurally blended clusters, with the example cluster described in Figure 9B exhibiting an RMSD of 3.92 between the two configurations (Supplementary Table 4), and how relatively common they are, it is increasingly clear that while structural information is indeed being recovered to a certain extent, it is also likely that some other factor correlates just as if not more strongly, revealing the possibility that other biological information being recovered. What remains to be understood is what this information could be and how it relates to the observed structural recovery we see. These patterns can also be seen in PCA derived configurations which, although hard to distinguish due to the point distribution density, can be identified when graphing a smaller number of configurations (Figure 8) or when mapping these same configurations onto the UMAP reduction graphs (Figures 6C & 7C). When considering the higher-than-expected RMSD values of these structurally blended clusters, with the example cluster described in Figure 9B exhibiting an RMSD of 3.92 between the two configurations (Supplementary Table 4), and how relatively common they are (Supplementary Table 4 & 5), it is increasingly clear that while structural information is indeed being recovered to a certain extent, it is also likely that some other factor correlates just as if not more strongly, revealing the possibility that other biological information being recovered. What remains to be understood is what this information could be and how it relates to the observed structural recovery we see. These patterns can also be seen in PCA derived configurations which, although hard to distinguish due to the point distribution density, can be identified when graphing a smaller number of configurations (Figure 8) or when mapping these same configurations onto the UMAP reduction graphs (Figures 6C & 7C).

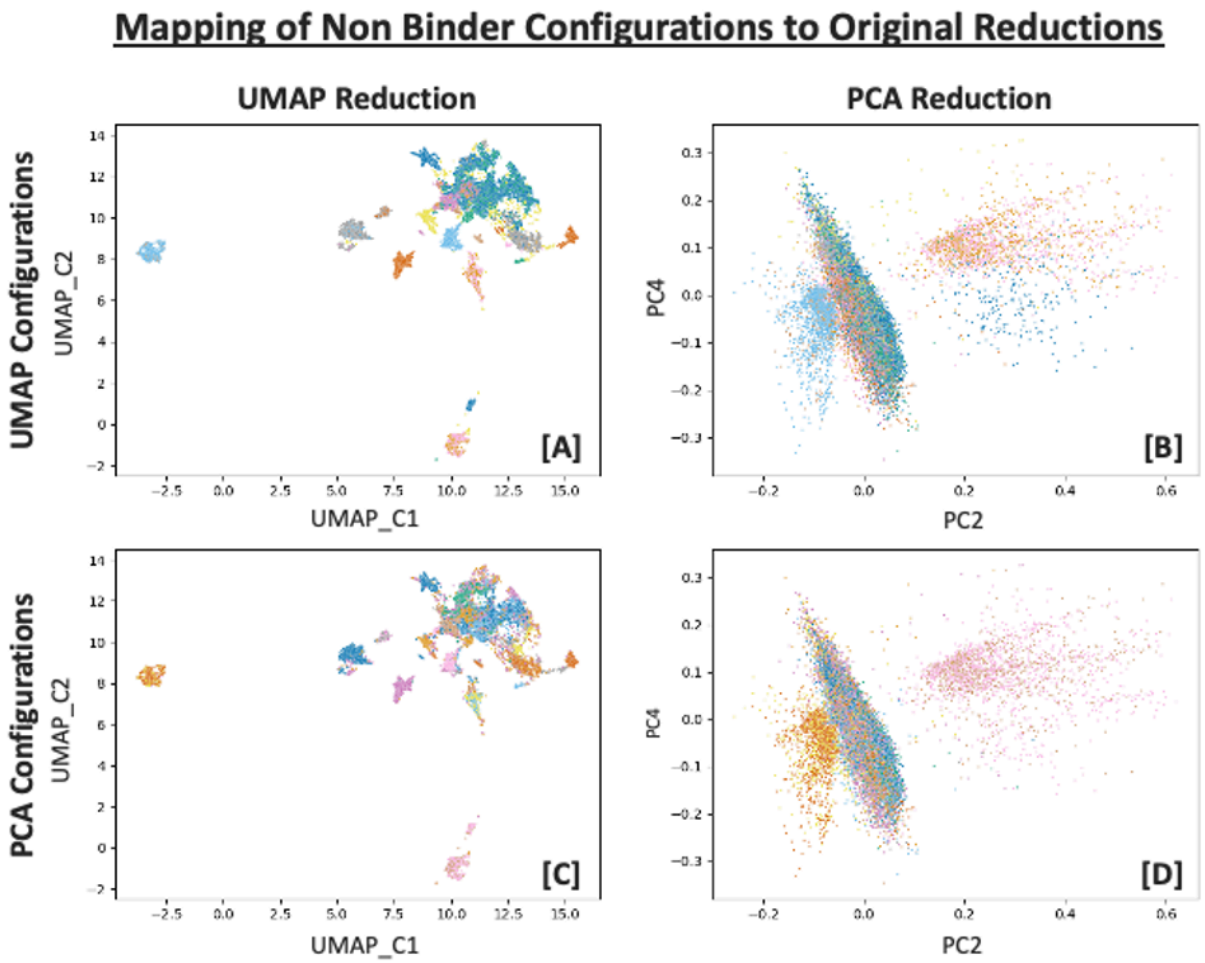


Figure 6. Mapping of non-binder structural configurations exhibit significant overlaps with starting reductions, though far harder to identify in PCA reductions. *Mapping of UMAP (A, B) and PCA (C, D) derived binder configurations onto original UMAP (A, C) and PCA (B, D) binder reductions (see Figure 1A/D for original point distributions and Figure 2B UMAP clusters). Results are broadly similar to results observed with binders (Figure 7), strong overlap between UMAP-derived configurations and UMAP reduction clusters can be identified (A), which, interestingly, can also be extended PCA-derived data mapped to UMAP (C), with seemingly finer structural details being revealed than UMAP to UMAP (A), particularly of the large central UMAP cluster (Figure 2B). Though some broadly patterns can be identified, such overlaps are harder to detect in PCA reductions (B, D) without additional filtering due to the point distribution (see Figure 8 for example of filtered visualization). See Supplemental Data for mappings on all tested reductions.*

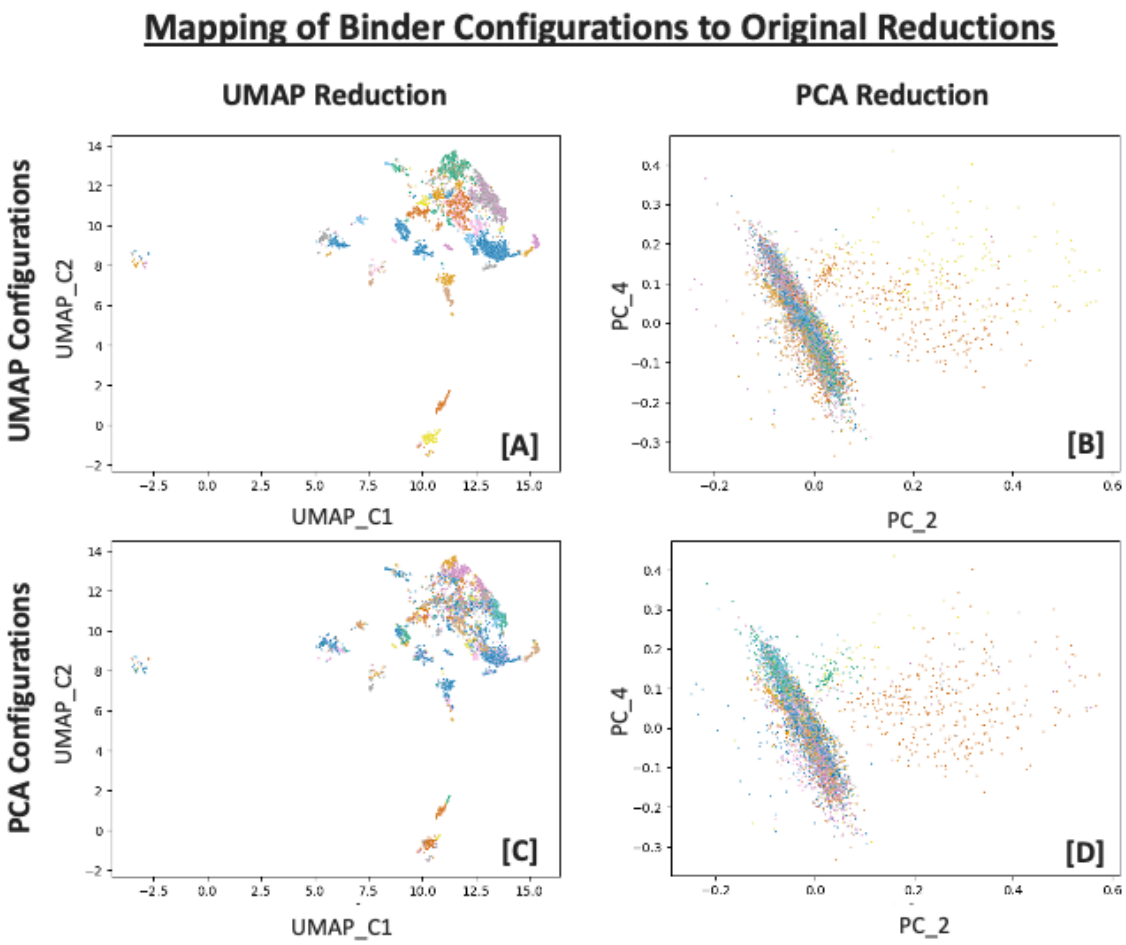


Figure 7. Mapping of binder structural configurations exhibit significant overlaps with starting reductions, though far harder to identify in PCA reductions. *Mapping of UMAP (A, B) and PCA (C, D) derived binder configurations onto original UMAP (A, C) and PCA (B, D) binder reductions (see Figure 1A/C for original point distributions and Figure 2A UMAP clusters). Results are broadly similar to results observed with nonbinders (Figure 6), with strong overlap between UMAP-derived configurations and UMAP reduction clusters can be identified (A), which, interestingly, can also be extended PCA-derived data mapped to UMAP (C), with seemingly finer structural details being revealed than UMAP to UMAP (A). Though some broadly patterns can be identified, such overlaps are harder to detect in PCA reductions (B, D) without additional filtering due to the point distribution (see Figure 8 for example of filtered visualization). See Supplemental Data for mappings on all tested reductions.*

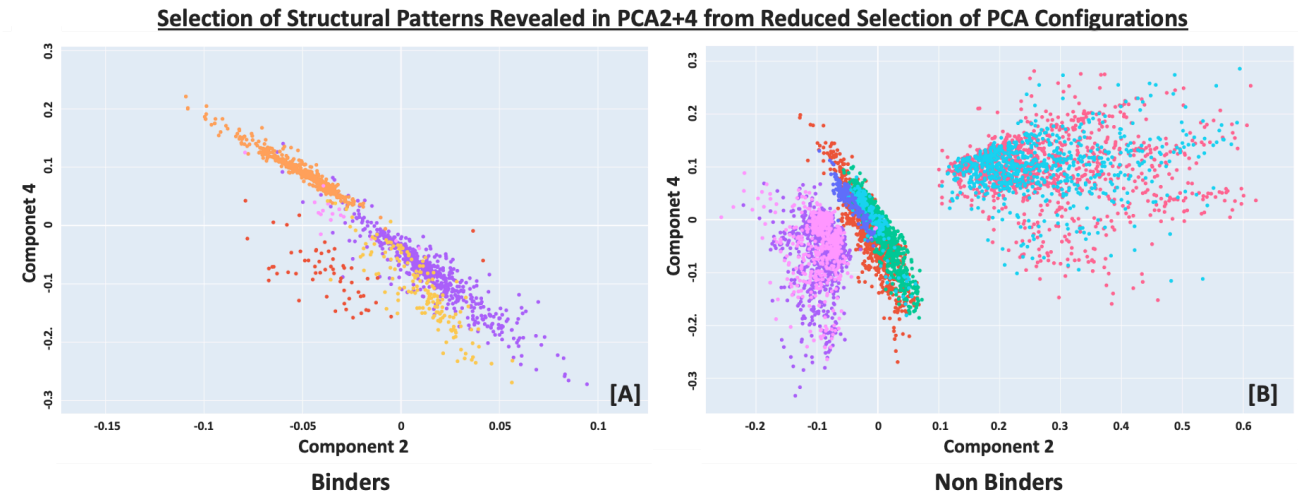


Figure 8. Selective screening of structural configurations in PCA reductions reveals structural patterns in distribution hidden by initial point density for both binders & nonbinders. *Mapping of a selection 5 binder (A) and 8 non-binder (B) PCA-derived configurations onto PCA components 2 & 4. Demonstrates that clear distribution patterns are occurring for both types, which are obscured when all points are analyzed simultaneously (compare to Figures 6C/D & 7C/D). Additional hindrance is caused by the configurations tracking fairly closely with the overall spike-like distribution of the starting graph, making pattern identification on the full dataset difficult.*

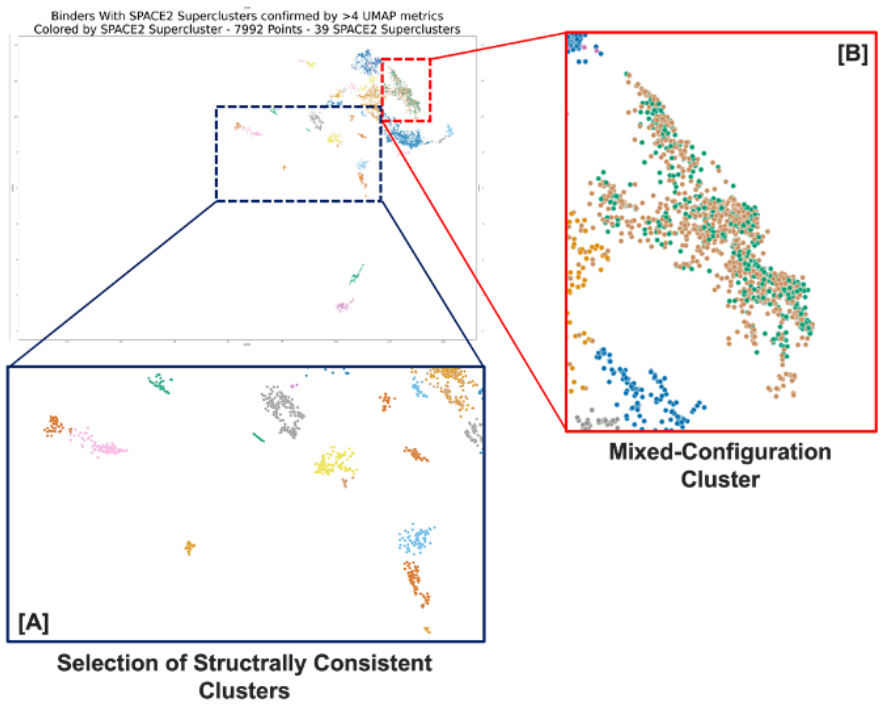


Figure 9. Larger UMAP clusters tend to exhibit a relatively uniform blending of two structural configurations with higher than expected RMSD values, while most smaller clusters are composed of a single overall configuration. *Focused images of a blended mixed-configuration (B) and structurally consistent (A) binder clusters mapped onto example UMAP reduction (see Figure 2A for original reduction clusters and 7A for full configuration mapping). Mixed-configuration cluster displays two colors (orange vs green) blended relatively equally, with an RMSD value calculated at 3.92, far higher than the set RMSD cutoff or the general binders-only RMSD average. Comparatively, clusters identified in (A) all consist of a single color, confirming the presence of a single configuration in each.*

3.4. Analysis of Cluster Residue and Physicochemical Diversity

3.4.1. Confirmation of Sequence Diversity in Reduction Clusters

The significant sequence heterogeneity observed within both single and multi-structural reduction clusters (Figure 10) corroborates that the model-based classification transcends classification by sequence similarity, further corroborating the initial full dataset sequence diversity observations made in section 2.1.

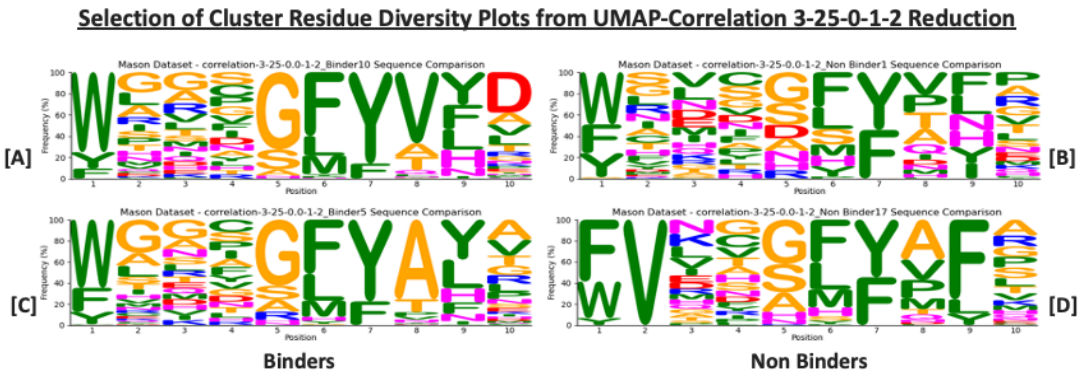


Figure 10. Both binder and non-binder clusters exhibit significant sequence diversity with biases similar to those observed in the full dataset. *Selection of diversity plots from selected example UMAP-correlation reduction demonstrating sequence diversity of binder (A, C) and non-binder (B, D). Minimal overall differences can be observed from full dataset analysis, with the broad residue and physicochemical variability patterns described in Figure 4 remaining consistent across clusters and label types. Number of sequences analyzed: (A)"Binder10": 1,772; (B)"NonBinder1": 12,493; (C)"Binder5": 970;(A) "NonBinder17": 119. See Supplemental Data for all generated logoplots.*

3.4.2. Broad Conservation of Library-Wide High & Low Variability Regions Across Clusters

Observations of residue diversity were congruent with broader library findings, echoing biases in diversity identified in the full dataset analysis (Figures 4 & 10). Low diversity positions 1, 6 and 7 seen in the full dataset analysis exhibit consistently reduced diversity in most reduction and structural clusters. Conversely, positions 2-4 maintain their relative diversity across most analyzed clusters. Some of these may exhibit cluster-specific changes in diversity ratios at specific positions (Figure 11, 12), however most the overall relative diversity levels seen in the full dataset analysis at these specific positions. Additionally, no broader patterns between binder and non-binder clusters can be identified at these specific positions, similar again to the results seen in the full sequence analysis.

Positions 8-10, on the other hand, while still generally showing similar levels of high diversity, reveal consistent patterns that can be identified across multiple reduction clusters. For example, a relatively frequent F9D10 motif can be observed in many binder clusters (Figure 11), while most reductions contain a non-binder cluster dominated by a lysine or arginine at position 8 (Figure 12), indicating that such motifs may be playing a role in sequence classification. While this may appear to be evidence of clustering based on sequence similarity, it is important to consider that these identifiable motifs are still quite rare considering the total number of clusters analyzed, with these positions still exhibiting an overall high level of diversity in the majority of other clusters.

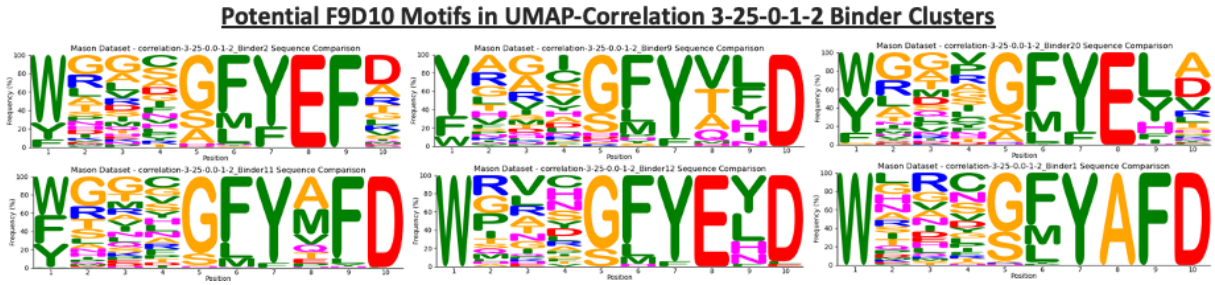


Figure 11. Some binder clusters appear to exhibit strong residue and/or physicochemical property distributions at certain positions. *Examples of a potential F9D10 motif identified at positions 8-10 in UMAP binder clusters. All sequence diversity plots are of reduction clusters isolated from the example UMAP-correlation graph, though such patterns were consistently identified across UMAP reductions. Given these recurring distribution patterns, these serve as possible evidence of the model's ability to recover some kind of small-nonpolar or negatively charged + hydrophobic + negatively charged residue motif, and thus some physicochemical properties more broadly. Number of sequences analyzed, clockwise from top left: "Binder2": 381; "Binder9": 95; "Binder20": 109; "Binder1": 86; "Binder12": 36; "Binder11": 193*

3.4.3. Potential Evidence of Physicochemical Patterns Within Clusters

The examination of the distribution of the residue physicochemical properties within and across clusters unveils tantalizing preliminary evidence suggesting a potential correlation with the clustering patterns observed. For instance, as mentioned 3.4.2, most reductions contain a non-binder cluster with a lysine being the exclusive residue at position 8 across all sequences within the cluster (Figure 12). This is just the most prominent evidence of a broader physicochemical pattern seen across all logo plots, with a notable absence of positively charged residues at position 8 in binder clusters (Figure 10 & 11). Similarly, the F9D10 motif described above may be better interpreted as a physicochemical motif of a hydrophobic residue sandwiched by two negative residues (Figure 10). While far from conclusive, these and other less substantiated observations are quite noticeable given the overall domination of hydrophobic and small-nonpolar residues across all cluster sequences and underscores a potential relationship with clustering outcomes. Such insights may point to the other kind of information recovered by the model being some type of physicochemical properties.

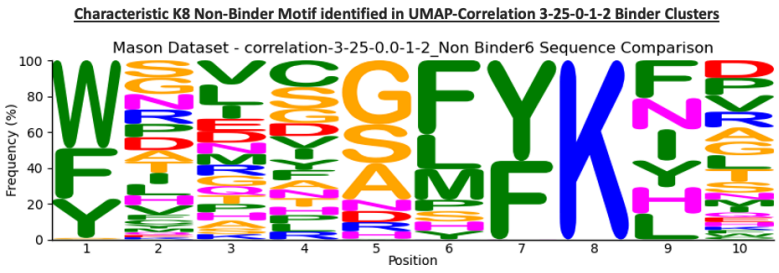


Figure 12. Potential evidence of physicochemical recovery in non-binder clusters. *Example of K8 motif seen in at least one non-binder cluster across most reductions. Exemplative of a strong presence of positive residues, particularly lysine, often being overrepresented at position 8 in nonbinders, compared to their relative absence in binder clusters (Figure 10).*
"NonBinder6": 1911

3.5. Insights into Dimensionality Reduction Approach

3.5.1. Minor Differences in Overall Efficiency of UMAP Distance Metrics

A comparative analysis of UMAP distance metrics revealed subtle variances in delineating structural relationships with the dataset, though quite minimal in most instances. Internal structural consistency, reduction cluster coverage and internal RMSD values (Figure 14, Supplementary Table 1), all metrics were broadly consistent with one another, exhibiting only minor differences in overall values. The main significant difference observed was in the number of sequence superclusters captured, providing indirect evidence in the ability of each metric to successfully group all sequences into recognizable clusters at least once given the variety of parameters tested.

Among the metrics assessed, correlation and cosine demonstrated superior and consistent efficacy on all standards tested, with correlation exhibiting a marginal advantage in sequence supercluster identification by covering 34138 sequences to cosines 34024 (Table 1). Manhattan and Euclidean based reduction displayed comparatively limited utility, with manhattan being particularly ineffective compared to other metrics (Figure 14), suggesting it should not be used in future research of this kind. Interestingly, the substitution based Hamming distance, while performing similarly to the euclidean and hamming metrics, demonstrated a substantial ability to pick up sequences that remained unidentified by both cosine and correlation reductions, despite the metric only identifying 33784 sequences (Table 1). In total, of the 9 sequences that remained unidentified using correlation distance, data from hamming reductions was able to identify 6, resulting in only 3 remaining unidentified sequences (Table 1). In comparison, cosine data successfully identifies only 4. While a minimal difference in the result, it is quite startling when

	Correlation	Cosine	Euclidean	Hamming	Manhattan	Correlation + Cosine	Correlation + Hamming	Correlation + Hamming + Cosine	All Metrics
Binder	8946	8832	8660	8592	8926	8950	8951	8952	8954
nonbinders	25191	25191	25191	25191	18457	25191	25191	25191	25191
Total	**34137**	**34023**	**33851**	**33783**	**27383**	**34142**	**34143**	**34144**	**34145**
Missing Binders	9	123	295	363	29	5	4	3	1
Missing NonBinders	0	0	0	0	6734	0	0	0	0
Total Missing	**9**	**123**	**295**	**363**	**29**	**4**	**3**	**2**	**1**

Table 1. All UMAP metrics except Manhattan display extremely high capabilities in capturing the vast majority of sequences in at least one reduction cluster. Number of sequences that each metric was able to successfully cluster, as well as correlation + cosine, correlation + hamming, correlation + hamming + cosine and across all metrics. Can observe that Manhattan was significantly less efficient in non-binder capture, while all other metrics successfully identified all nonbinders. In total, only one sequence remained unclustered across all metrics, while combining correlation with cosine and/or hamming significantly reduced the amount of unidentified binders

considering that hamming-derived data identified 244 less sequences than the cosine data, underscoring the differences in clustering between the two distance methods and the potential usefulness in using hamming as a layer of redundancy for maximum sequence capture.

3.5.2. Effectiveness of Combined Correlation & Hamming Distance Metric Use

As discussed in section 3.5.3, the high efficiency in identifying a single structural supercluster for the vast majority of sequences by any distance, but particularly correlation and cosine, underscores the robust structural characterization capacity of this methodology. The minimal differences in supercluster identification and overall structural configurations of the reduction clusters suggest that an exhaustive optimization using all metrics is not imperative (Table 1). In consideration of a streamlined approach, the ability for a combination of correlation and hamming distance alone to result in only 4 unclustered sequences across all reductions suggests a potentially optimized methodological framework for future research (Table 1). Unfortunately, little else can be concluded from the broader patterns between the UMAP metrics as all exhibited similar configuration mapping with minimal observable differences, indicating that some degree of optimization will most likely be required in future research until broader patterns can be determined, if at all possible.

3.5.3. Difficulties in Interpretation of PCA Derived Data

As can be intuited from the overall lack of analysis of PCA results in previous sections, PCA-based point distribution, while yielding some interesting structural patterns and results, proved to be difficult to analyze to to our inability to produce the small, focused clusters this study aimed for (Figure 13), particularly in comparison to the distinct cluster derived via UMAP (Figure 5). This can be attributed to the aggregation of data points into a handful of pronounced dense spikes (Figure 1B) which complicate data analysis, interpretation and conclusion. The structural superclusters were consistent across all PCA components, resulting in a single identified structural configuration for each sequence, which, when combined with the overall lack of significant separation between labels,

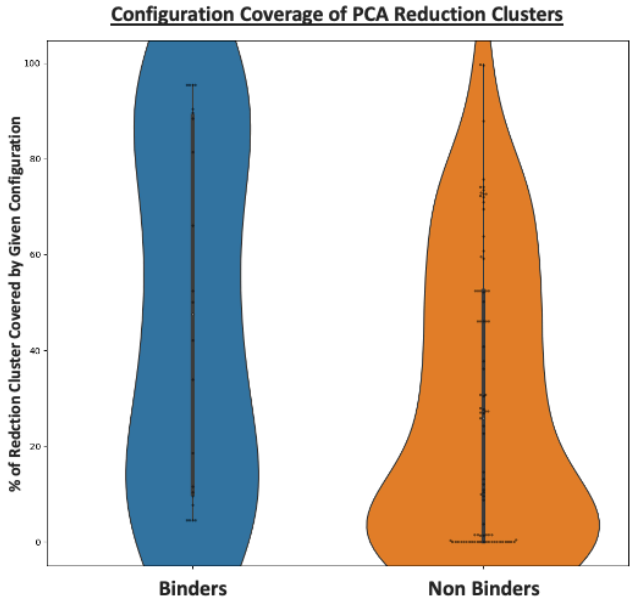


Figure 13. Manually extracted PCA clusters exhibit far less structural consistency due to their large size. *The percent amount of an associated reduction cluster a given configuration represents, for all reduction clusters across all PCA reductions. 23 binder configurations (blue) analyzed with mean of 47.82 ± 36.99% & median of 47.60%. Compared to 87 binder configurations (orange) analyzed with mean of 28.74 ± 29.27% & median of 25.90%. Distribution demonstrates the difficulty in the use of this manual cluster methodology in successfully identifying correlations between reduction clusters and structural configurations due to the reduction clusters extracted being far too broad.*

imply that additional exploration of higher PCA components may not yield substantially new insights. Upon close analysis, some patterns in the distribution of structural configurations can be observed, however, in aggregate, most of the manually isolated clusters did not exhibit the same strong correlation with structural configuration. However, this does not necessarily indicate a broader inability of PCA techniques to identify structural patterns since some can still be identified (Figures 6C, 7C & 8), simply that the methodology tested may be insufficient in capturing more specific structural patterns. Additionally, the unique distribution morphology observed in the PCA reductions may also warrant future exploration, as the distinct spike-like patterns could potentially prove useful in elucidating trends governed by non-categorical variables or in position specific analyses.

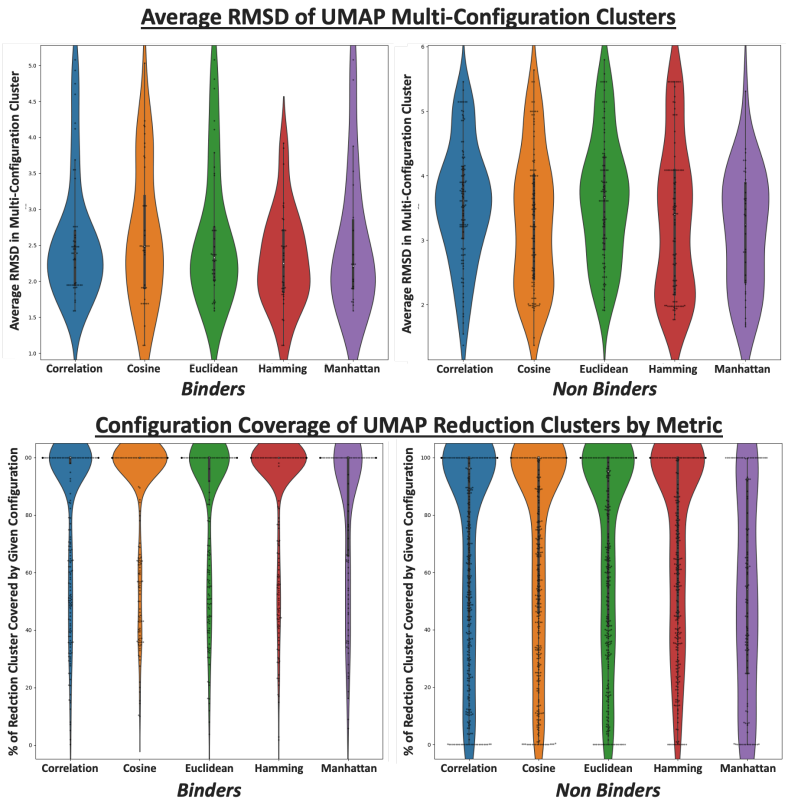


Figure 14. UMAP Manhattan distance displays the most significant lack of overall efficiency compared to the broad similarities seen across all other distance metrics. *Visualization of configuration percentage of associated reduction clusters (bottom) and average RMSD values of multi-configurational clusters (top) by metric and label. Can observe that all metrics are broadly comparable in both measures for both labels with some variations, particularly in non-binder RMSD values. Most significant observation is the consistently poorer results seen when using Manhattan distance, particularly in non-binder coverage & RMSD values. See Supplementary Table 1 for statistics of each metric.*

4. Discussion

4.1. Summary of Results

4.1.1. Clustering Patterns Indicate Model's Classification Transcends Raw Sequence Data

Systematic analysis across the full dataset reveals consistent residue diversity bias as expected from experimentally derived libraries. These patterns in sequence and physicochemical diversity are broadly mirrored within the reduction clusters, indicating that clustering is not based solely on raw sequence similarity. While some clusters can exhibit residue specific biases, with a single residue at a given

position for a given cluster, the overall prevalence of clusters exhibiting little to no sequence bias indicates that raw sequence is not the primary method by which clustering is occurring.

4.1.2. Successful Identification Structural Patterns Through Integrated Gradient Analysis

The recovery of some amount of structural information was confirmed by the UMAP reduction clusters exhibiting significant structural consistencies, particularly by the number of binder clusters composed of singular structural configurations. Non-binder clusters, while exhibiting less of this singular structural configuration pattern, still displayed considerable structural consistency, particularly for clusters distributed further from the central core of the reduction. This suggests that, while some structural information is being recovered, the degree of structure recovery by the model is ultimately different based on a sequence's ultimate classification as a binder or non-binder.

4.1.3. Inconsistencies in Structural Recovery Indicative of Differential Classification

Variability within some reductions clusters further points to complex interactions between sequence structural attributes and the model's clustering process, with differences in RMSD distributions between binder and non-binder mixed-configuration clusters further cementing this differential treatment by the model or clustering methodology. However, the significant RMSD ranges exhibited in the mixed-configuration binder clusters indicate that such interactions are not limited to nonbinders.

4.1.4. Possible Evidence of Clustering Based on Physicochemical Properties

Though falling outside the scope of this study's research question, analysis of physicochemical property distributions indicate the possible recovery of physicochemical properties by the model in addition to the observed overlaps with structural configurations. Notable motifs and residue patterns, such as the F9D10 motif in specific binder clusters and K8 motif in nonbinders, suggest a correlation between these properties and clustering outcomes. When considering the larger-than-expected RMSD values of certain blended mixed-configuration reduction clusters, these patterns become particularly interesting as indications of these different types of data recovery and potential jumping off points to study their relationship in future research projects.

4.2. Potential Effects of Unconsidered Factors On Dimensionality Reduction Optimization

The challenges encountered in the optimization of the UMAP and PCA reductions underscore the intricacies related to the selection of optimal parameters for analyzing such high-dimensional IG biological data. In particular, this methodology emphasizes the need for careful tuning combined with computational efficiency to achieve meaningful results. While broadly similar, the slight increase in structural consistencies and sequence identification observed using correlation distance, particularly when

coupled with hamming distance as added redundancy, confirms its use as the primary analysis method for future research using such data.

While all reductions methods and UMAP distance methods showed significant effectiveness in identifying sequence-specific membership to structural clusters, slight differences in overall efficiencies and result interpretability lead to several observations and conclusions. Regarding UMAP reductions, the use of correlation and hamming showed the most significant combined effectiveness and redundancy potential, suggesting their optimized future use to enhance efficiency while maximizing sequence coverage.

Conversely, although PCA reductions are capable of detected structural patterns congruent with those observed in UMAP, the point distributions observed provide significant challenges in analysis and interpretation, particularly in determining what role recovered structural information may play in classification. However, given that the PCA derived structural configurations overlap quite significantly with the UMAP reduction clusters, and that some distinct structural clustering can be seen to occur within the PCA graphs when carefully filtered, it is clear that some information is being recovered, with the main challenge arising from identification and interpretation. Interestingly, it should be noted that PCA distribution using the previously derived Absolut! synthetic dataset displayed a point distribution that was much more similar to the UMAP reductions optimized for this study. This could potentially be a byproduct of the inherent biases demonstrated in the experimental dataset used, potentially providing additional clues regarding the model's ability to recover biologically relevant data. When considering the potential physicochemical patterns observed, the distinct spike-like distribution patterns could prove more useful in analysis of the quantitative properties of residues and sequences, such as isoelectric point or RMSD evolution from a central point.

Ultimately, these results are indicative of only a single experimental library of CDRH3s for a single antigen. When considering the differences in distribution seen between the PCA of experimental vs a synthetic dataset, it remains to be seen if these distributions and therefore proposed methodologies remain consistent across different antigens, CDR sequences and experimental methodologies. Future studies should therefore be aware of how such differences affect point distribution and which, if any, of these factors play a determinant role in providing optimal reductions using either method.

4.3. Limitations in Structural Analysis & Possible Effects on Observed Inconsistencies

The identification of a single structural configuration across virtually all sequences as well as the consistent observed overlap across a variety of reduction clusters suggest the potential of such deep neural network models to yield biological relevant information beyond the simple classification task for which they are designed. However, the consistent presence of structural incongruities, such as the blended mixed configuration clusters and the strong differences between binder and non-binder clusters, prompts the consideration of the recovery of additional non-structural factors playing at least a semi-determinant role. Of particular interest would be their relationship to the observed structural recovery.

However, it should be mentioned that this study, by focusing on the CDRH3 structural consistency within selected binder or non-binder clusters, did not consider the possibility of using such structural clustering and analysis of the full sequence library. While the confirmation of relative structural consistency within individual clusters is significant, it remains to be determined if a full library structural analysis could yield similar overlaps of structural configurations with reduction patterns. A future exploration attempting to replicate these observed patterns would prove quite useful in resolving the significant computational and workflow inefficiencies inherent to the methodology described in this study.

That isn't to say that the efforts presented above were misguided, rather that the time constraints and desired flexibility in analysis made this the best approach given this study's objectives. Now that these structural patterns have been identified, the ability to replicate them by clustering the full library using SPACE2 or some other structural clustering algorithm could provide significant computation and time savings, particularly by shortening or fully eliminating the reduction cluster and calculation steps. However, such efforts will most likely require significant optimizations and explorations of both the SPACE2 and other structural clustering algorithms, with particular focus on the effects of decreasing the RMSD cutoff to see if more specific structural patterns can be identified. Despite these limitations, the presence of structural patterns identified within the large PCA clusters provides encouraging evidence that such an approach could yield successful results in the future.

Similarly, it is important to note that these results were derived from an artificial segregation of binder and non-binder sequences during the entirety of the analysis. While this analysis relied on reductions that showed maximum separation between the two classes, particularly for UMAP, the degree to which this separation occurred could vary significantly, as seen in the PCA reductions, thus is not a pattern that we should immediately assume for all future datasets. Additionally, even in the best of circumstances, some degree of mixing between the two labels is always present, which would have resulted in some of the opposing labels being present in clusters dominated by the other type without this artificial segregation. Furthermore, this approach resulted in the structural clustering also being segregated, with binders structures only being clustered with other binders and vice versa. While this proved fortuitous within this study, allowing for the identification of differences in structural configurations within label-specific clusters that would have otherwise been muddied, it does not allow for the overall configuration distribution within the reductions and the resulting conclusions. For example, this analysis cannot determine if sequences with a given label share structural similarities to sequences of the opposing label contained within the same overall cluster. Such an exploration could prove quite fruitful however, potentially revealing further information on the observed data recovery.

4.4. Preliminary Recovery of Physicochemical Patterns Provide Basis for Further Explorations

The observations of high sequence diversity within the analyzed data confirm that the model transcends reliance on raw sequence data in its classification calculations, rather than as a consequence of CDRH3s being clustered together due to sequence similarity, with resulting structural similarities occurring as a result of that. While some reduction clusters do exhibit dominance of certain residues at specific

positions, their relative rarity within and across reductions indicate that the raw sequence itself is most likely not a causal factor in this clustering effect.

Most interestingly however, this shallow exploration has revealed interesting patterns in the distribution of residue physicochemical properties, suggesting a potential correlation with clustering outcomes, which in turn indicates that such properties may also be recovered and play an integral role in the model's classification outcomes. This possibility is particularly relevant given the inconsistencies discussed, such as the high occurrence of blended mixed-configuration binder clusters with significantly elevated RMSD values or the relative non-binder structural chaos found within the core of the reductions.

These inconsistencies are both a limitation to the establishment of any definitive conclusions regarding the model's interpretation of recovered structural information and an opportunity to consider if the results seen in this physicochemical property analysis could be a potential explanation for said inconsistencies by pointing to the recovery of both physicochemical and structural information. Within this specific dataset, the dominance of small-nonpolar and hydrophobic residues throughout the library and clusters makes the relative dominance of positive, negative and hydrophilic residues within specific clusters of particular interest. As noted in previous sections, the distribution of these charged residues can rise to the point of potentially identifiable motifs or patterns, such as the hydrophobic/positive residues at position 9 & 10 in binders, or the strong concentration of positive residues at position 8 in nonbinders.

However, these findings are extremely preliminary and conditional on deeper and far more focused studies to understand such relationships and how they correlate to the structural inconsistencies observed. The results of such detailed analyses could prove critical in not only our understanding of the operations of such models, but also with how such recovered biological information can be best used towards future research and practical applications. For example, it remains to be determined whether the structural congruences observed are concomitant byproducts of these potential physicochemical properties, being preferentially clustered together due to their high correlation with structural similarities. While it is still not clear what the best methods to explore these potential relationships would be, the final section of this study will focus on proposing several starting points with the aim to begin elucidating these questions.

4.5. Structural Recovery Opens New Possibilities in Sequence Based Antibody Modeling

Overall, despite the limitations discussed in the previous sections, the demonstrated ability of these models to extract biologically relevant information solely on sequence data underscores their potential utility in future research. Recent approaches often rely on models that utilize antibody structures as inputs, which, while effective, can be computationally intensive and time-consuming by requiring the computation of the antibody structure. The findings from this study suggest that using sequence-based could significantly streamline these processes by avoiding the intensive task of structural modeling while still recovering structural or other types of information, opening up up new possibilities for rapid and extensive exploration of CDR and broader antibody repertoires and leading to new insights into antibody behavior and functionality.

5. Future Work

5.1. Optimization of SPACE2 Structural Clustering

A pivotal focus of future studies will be on the proper optimization and employment of the SPACE2 structural clustering algorithm, combined with in depth analysis on how these optimization efforts affect the distribution of structural configurations within selected reductions. In that regard, a simple first step would be the clustering of the full, binders-only and nonbinders-only datasets using the default SPACE2 parameters and compare with the structural distribution patterns observed in the single reduction cluster analysis conducted in this study. Based on the results seen in both the UMAP and PCA reductions, where distinct configurations can be identified within individual reduction clusters, confirmation of the emergence of these patterns independent of reduction cluster isolation could prove beneficial in maximizing workflow efficiency by removing the need to isolate sequences into separate clusters prior to structural analysis.

Subsequent work should then be conducted to explore how changes in SPACE2 parameters affect configuration distribution within the existing reductions. As mentioned previously, results of this study were obtained using the recommended default parameters of SPACE2 to analyze only the CDRH3 structures. For example, analyzing the resulting differences when clustering both CDRH3 and the full IgG structures could provide valuable insight in the best use to identify the strongest patterns of structural configurations. Additionally, an in depth exploration of the effects of changes in the RMSD cutoff parameter, especially by setting it at the average binders-only RMSD value, would perhaps prove most interesting given the inconsistencies observed. That is not to discount the other parameters however, as an understanding of the changes that occur by setting different n_jobs or algorithms could also provide a deeper understanding of these distributions while at the very least optimizing our use of such software.

Such experiments should then be extended to alternative structural clustering models such as FOLDSEEK as a way to compare results and assess their effectiveness in retrieval of biologically relevant information relative to SPACE2. While significant, the proper implementation of these steps would allow for a better understanding of this methodology and these kinds of DNN models, allowing for the development of extremely refined workflows while allowing researchers a better understanding of the biological information being recovered and its best application in future studies.

5.2. Confirming Methodology Application and Reduction Optimization Across Diverse Datasets

While the results observed in this study are promising, future studies will first have to consider if the application of the methodologies described above can be extended to a broader range of datasets, both synthetic and experimental, and across different antigens. This measure will be critical in determining how the characteristics of different datasets influence the distributions observed in the optimized reductions. This work should include a further fine-tuning of the UMAP and cluster extraction parameters

described in this study to identify the optimal number of components and reductions needed to successfully cluster the majority of sequences to maximize this methodologies efficiency and utility.

5.3. In-Depth Exploration of Physicochemical Property Recovery

The observations of a potential correlation between physicochemical properties and reduction point distributions indicate that future studies should expand focus to include analyses of categorical and quantitative representations of these properties in addition to the further investigations into recovered structural information. Such potential factors could be representations that track the evolution of RMSD values, sequence isoelectric points, or residue charge at individual positions. To that end, the distinct spike-like distributions seen in the PCA reduction could prove particularly interesting, as the focus of points along a handful of axes could allow for easier analysis and interpretation of results compared to the distributions seen in the UMAP reductions.

Moreover, special attention should be focused on the in-depth analysis of mixed-configuration reduction clusters to delve deeper into the complex relationship between the overlap in IG clustering, recovered structural information and physicochemical similarities. Of particular interest would be in determining if the structural data is being recovered independently of any potentially recovered physicochemical data, or if the structural overlaps seen are simply a byproduct of sequences sharing some hidden physicochemical attributes, which are correlated with increased overall structural similarity. Furthering our understanding of these relationships would allow for the expansion in the use of such classification models and their recovered biological information, opening a plethora of new research avenues and applications in the field of in-silico antibody research and design. One potential methodological approach could be a comparison of sequence alignment to the generated Lesk colored logo plots to confirm the validity of the motifs identified.

5.4. Application of Methodology Using Different Feature Attribution Methods

While the DR in this study analyzed IG values, other feature attribution methods may offer further nuanced insights. For example, XRAI, which highlights overlapping regions rather than individual positions to create saliency maps[33] , could enhance the identification of broader structural motifs across entire IgG constructs, while Sampled Shapley may elucidate epitope-specific physicochemical biases through its evaluation of different feature permutations[26,27]. An in depth exploration of these methods could provide critical optimization insights, potentially enabling more comprehensive recovery of various biological data types through their simultaneous application.

6. Ethical Reflections

On the promise of replacing experimental datasets through the simulation of large-scale synthetic antibody datasets, the further development of the results found in this study would hold considerable

benefits, in particular through reductions in laboratory resources, its associated single-plastics use, and shipping-related pollution. However, perhaps one of the crucial ethical advantages would be the ability to significantly decrease the field's reliance on animal models, particularly in hybridoma libraries which require animal sacrifice to generate. This study specifically is an example of this, where the ability to reuse hybridoma cell-line data derived in a previous study demonstrates the potential of utilizing existing data without additional animal sacrifices. However, careful attention should be paid during the development of such workflows to increase computational power consumption and its associated e-waste, and the proper disposal of the latter. Overall, given the potential of the long-term aims of this study and the overall simple moral considerations, this study and associated research direction is ethically justifiable and would see minimal opposition in being pushed further.

References

1. Schmitz S, Schmitz EA, Crowe JE Jr, Meiler J. The human antibody sequence space and structural design of the V, J regions, and CDRH3 with Rosetta. MAbs. 2022 Jan-Dec;14(1):2068212.

2. Lu RM, Hwang YC, Liu IJ, Lee CC, Tsai HZ, Li HJ, et al. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci. 2020 Jan 2;27(1):1.

3. Elsner RA, Shlomchik MJ. Germinal Center and Extrafollicular B Cell Responses in Vaccination, Immunity, and Autoimmunity. Immunity. 2020 Dec 15;53(6):1136–50.

4. Xu JL, Davis MM. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. Immunity. 2000 Jul;13(1):37–45.

5. Sivalingam GN, Shepherd AJ. An analysis of B-cell epitope discontinuity. Mol Immunol. 2012 Jul;51(3-4):304–9.

6. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody database. Bioinformatics. 2021 May 5;37(5):734–5.

7. Ferdous S, Martin ACR. AbDb: antibody structure database—a database of PDB-derived antibody structures. Database . 2018 Apr 27;2018:bay040.

8. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. Nucleic Acids Res. 2014 Jan;42(Database issue):D1140–6.

9. Akbar R, Bashour H, Rawat P, Robert PA, Smorodina E, Cotet TS, et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. MAbs. 2022 Jan-Dec;14(1):2008790.

10. Wardemann H, Busse CE. Novel Approaches to Analyze Immunoglobulin Repertoires. Trends Immunol. 2017 Jul;38(7):471–82.

11. Shiakolas AR, Kramer KJ, Johnson NV, Wall SC, Suryadevara N, Wrapp D, et al. Efficient discovery of SARS-CoV-2-neutralizing antibodies via B cell receptor sequencing and ligand blocking. Nat Biotechnol. 2022 Aug;40(8):1270–5.

12. Laustsen AH, Greiff V, Karatt-Vellatt A, Muyldermans S, Jenkins TP. Animal Immunization, in Vitro

Display Technologies, and Machine Learning for Antibody Discovery. Trends Biotechnol. 2021 Dec;39(12):1263–73.

13. Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. Cell Rep. 2021 Mar 16;34(11):108856.

14. Greiff V, Yaari G, Cowell LG. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. Current Opinion in Systems Biology. 2020 Dec 1;24:109–19.

15. Del Vecchio A, Deac A, Liò P, Veličković P. Neural message passing for joint paratope-epitope prediction [Internet]. arXiv [q-bio.QM]. 2021. Available from: http://arxiv.org/abs/2106.00757

16. Xu J. Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A. 2019 Aug 20;116(34):16856–65.

17. AlQuraishi M. End-to-End Differentiable Learning of Protein Structure. Cell Syst. 2019 Apr 24;8(4):292–301.e3.

18. Sverrisson F, Feydy J, Correia BE, Bronstein MM. Fast end-to-end learning on protein surfaces [Internet]. bioRxiv. bioRxiv; 2020. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.12.28.424589

19. Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S. Advancing Drug Discovery via Artificial Intelligence. Trends Pharmacol Sci. 2019 Oct;40(10):801.

20. Narayanan H, Dingfelder F, Butté A, Lorenzen N, Sokolov M, Arosio P. Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. Trends Pharmacol Sci. 2021 Mar;42(3):151–65.

21. Townshend RJL, Bedi R, Suriana P, Dror R. End-to-end learning on 3D protein structure for interface prediction. Adv Neural Inf Process Syst. 2018 Jul 3;15616–25.

22. Prakash E, Shrikumar A, Kundaje A. Towards more realistic simulated datasets for benchmarking deep learning models in regulatory genomics [Internet]. bioRxiv. 2021. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.12.26.474224

23. Cao Y, Yang P, Yang JYH. A benchmark study of simulation methods for single-cell RNA sequencing data. Nat Commun. 2021 Nov 25;12(1):6911.

24. Schuler A, Jung K, Tibshirani R, Hastie T, Shah N. Synth-Validation: Selecting the Best Causal Inference Method for a Given Dataset [Internet]. arXiv [stat.ML]. 2017. Available from: http://arxiv.org/abs/1711.00083

25. Robert PA, Akbar R, Frank R, Pavlović M, Widrich M, Snapkov I, et al. Unconstrained generation of synthetic antibody-antigen structures to guide machine learning methodology for antibody specificity prediction. Nat Comput Sci. 2022 Dec;2(12):845–65.

26. Maleki S, Tran-Thanh L, Hines G, Rahwan T, Rogers A. Bounding the Estimation Error of Sampling-based Shapley Value Approximation [Internet]. arXiv [cs.GT]. 2013. Available from: http://arxiv.org/abs/1306.4265

27. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks [Internet]. arXiv [cs.LG]. 2017. Available from: http://arxiv.org/abs/1703.01365

28. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng SM, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat Biomed Eng. 2021 Jun;5(6):600–12.

29. Ruffolo JA, Chu LS, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun. 2023 Apr 25;14(1):2389.

30. Lesk A. Introduction to Bioinformatics. OUP Oxford; 2014. 371 p.

31. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD [Internet]. 1996; Available from: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf?source=post_page----------

32. Spoendlin FC, Abanades B, Raybould MIJ, Wong WK, Georges G, Deane CM. Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind to the same epitope. Front Mol Biosci. 2023 Sep 18;10:1237621.

33. Kapishnikov A, Bolukbasi T, Viégas F, Terry M. XRAI: Better Attributions Through Regions [Internet]. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1906.02825

Appendices

	Correlation		Cosine		Euclidean		Hamming		Manhattan	
	Binder	Non-Binder	Binder	Non-Binder	Binder	Non-Binder	Binder	Non-Binder	Binder	Non-Binder
Clusters #	447	418	437	476	297	490	134	115	425	405
Average coverage	90.83%	71.53%	91.30%	71.43%	87.21%	72.65%	83.58%	53.04%	87.29%	72.84%
Coverage Std. Dev.	21.00%	34.65%	20.77%	35.21%	24.13%	34.44%	35.94%	35.94%	24.53%	32.45%
Median Coverage	100%	95.25%	100%	96.35%	100%	100%	100%	54.20%	100%	72.84%
Multi-Config Clusters #	40	96	38	107	38	108	22	37	54	99
Average RMSD	2.63	3.64	2.3	3.58	2.63	3.32	2.53	3.15	2.39	3.32
RMSD Std. Dev.	0.97	0.98	0.67	0.93	0.95	1.05	1	0.93	0.87	1.17
Median RMSD	2.49	3.66	2.25	3.61	2.33	3.32	2.21	3.25	2.39	3.4

Supplementary Table 1. Distance metrics are broadly consistent in terms of general structural consistency *Statistics from Figure 14, tracking median, mean and standard deviation of configuration percentage of associated reduction clusters and average RMSD values of multi-configurational clusters. Can identify broad consistencies across all metrics*

Cont.

Binder Structural Configurations (Excluding Out-of-Bounds Reduction Clusters)				
Structural Configuration	***Reduction Cluster***	***Priority***	***# of CDRH3***	***Reduction Cluster Coverage***
AgPos_fv_8712	***Binder0***	***high***	***1690***	***100***
AgPos_fv_6842	***Binder10***	***high***	***1135***	***64.1***
AgPos_fv_7785	***Binder4***	***high***	***773***	***81.5***
AgPos_fv_1450	***Binder5***	***high***	***682***	***70.3***
AgPos_fv_5689	***Binder10***	***high***	***637***	***35.9***
AgPos_fv_6692	***Binder14***	***med***	***318***	***100***
AgPos_fv_8503	***Binder5***	***high***	***288***	***29.7***
AgPos_fv_7095	***Binder2***	***med***	***261***	***100***
AgPos_fv_1295	***Binder7***	***med***	***204***	***100***
AgPos_fv_2833	***Binder16***	***med***	***195***	***100***
AgPos_fv_8894	***Binder11***	***med***	***193***	***100***
AgPos_fv_8876	***Binder22***	***med***	***180***	***100***
AgPos_fv_8322	***Binder4***	***high***	***176***	***18.5***
AgPos_fv_5690	***Binder8***	***med***	***152***	***100***
AgPos_fv_1399	***Binder26***	***med***	***116***	***100***
AgPos_fv_5828	***Binder20***	***med***	***109***	***100***
AgPos_fv_5492	***Binder31***	***med***	***104***	***100***

AgPos_fv_5073	*Binder27*	*med*	*101*	*100*
AgPos_fv_7203	*Binder6*	*med*	*97*	*100*
AgPos_fv_7701	*Binder9*	*med*	*94*	*100*
AgPos_fv_7043	*Binder19*	*med*	*93*	*100*
AgPos_fv_6622	*Binder23*	*low*	*88*	*100*
AgPos_fv_2269	*Binder1*	*low*	*86*	*100*
AgPos_fv_6358	*Binder3*	*low*	*86*	*100*
AgPos_fv_6629	*Binder25*	*low*	*78*	*100*
AgPos_fv_7649	*Binder13*	*low*	*64*	*100*
AgPos_fv_2516	*Binder24*	*low*	*59*	*100*
AgPos_fv_1555	*Binder17*	*low*	*53*	*100*
AgPos_fv_6019	*Binder28*	*low*	*50*	*100*
AgPos_fv_2788	*Binder29*	*low*	*45*	*100*
AgPos_fv_8884	*Binder34*	*low*	*44*	*100*
AgPos_fv_3905	*Binder15*	*low*	*38*	*100*
AgPos_fv_252	*Binder12*	*low*	*36*	*100*
AgPos_fv_2708	*Binder18*	*low*	*34*	*100*
AgPos_fv_1736	*Binder21*	*low*	*27*	*100*
AgPos_fv_5949	*Binder33*	*low*	*27*	*100*
AgPos_fv_5227	*Binder32*	*low*	*23*	*100*
AgPos_fv_6538	*Binder30*	*low*	*20*	*100*
AgPos_fv_3032	*Binder35*	*low*	*16*	*100*

Supplementary Table 2. Binder configurations exhibit strong correlations with reduction clusters.
Identified SPACE2 configurations in example UMAP-correlation with their associated reduction clusters. Contains the number of CDRH3 described by each configuration and the percentage of the associated reduction cluster it represents. Ordered from high to low by number of CDRH3 in each configuration.See Supplementary Table 3 for nonbinders

Non Binder Structural Configurations (Excluding Out-of-Bounds Reduction Clusters)

Structural Configuration	*Reduction Cluster*	*Priority*	*# of CDRH3*	*Reduction Cluster Coverage*
AgNeg_fv_22937	*Non Binder6*	*med*	*1648*	*86.3*
AgNeg_fv_12944	*Non Binder13*	*med*	*1570*	*100*
AgNeg_fv_21056	*Non Binder4*	*med*	*1479*	*82.7*
AgNeg_fv_23986	*Non Binder3*	*med*	*964*	*57*
AgNeg_fv_25008	*Non Binder10*	*med*	*839*	*100*
AgNeg_fv_12740	*Non Binder0*	*med*	*832*	*100*
AgNeg_fv_23527	*Non Binder2*	*med*	*747*	*100*
AgNeg_fv_10429	*Non Binder3*	*med*	*726*	*43*
AgNeg_fv_14372	*Non Binder8*	*med*	*612*	*58.3*
AgNeg_fv_5702	*Non Binder5*	*med*	*513*	*100*
AgNeg_fv_22430	*Non Binder8*	*med*	*436*	*41.5*
AgNeg_fv_21383	*Non Binder7*	*med*	*321*	*62.2*
AgNeg_fv_14826	*Non Binder4*	*med*	*309*	*17.3*
AgNeg_fv_17977	*Non Binder9*	*med*	*269*	*85.1*
AgNeg_fv_7798	*Non Binder6*	*med*	*262*	*13.7*
AgNeg_fv_14079	*Non Binder-1*	*low*	*238*	*100*
AgNeg_fv_19892	*Non Binder11*	*low*	*231*	*100*
AgNeg_fv_5789	*Non Binder7*	*med*	*195*	*37.8*
AgNeg_fv_3718	*Non Binder16*	*low*	*160*	*100*
AgNeg_fv_7894	*Non Binder17*	*low*	*119*	*100*
AgNeg_fv_1534	*Non Binder12*	*low*	*78*	*100*
AgNeg_fv_9302	*Non Binder9*	*med*	*47*	*14.9*
AgNeg_fv_7729	*Non Binder18*	*low*	*29*	*100*
AgNeg_fv_24389	*Non Binder14*	*low*	*28*	*100*
AgNeg_fv_1255	*Non Binder15*	*low*	*24*	*100*
AgNeg_fv_15500	*Non Binder19*	*low*	*20*	*100*
AgNeg_fv_635	*Non Binder8*	*med*	*1*	*0.1*
AgNeg_fv_16600	*Non Binder8*	*med*	*1*	*0.1*

Supplementary Table 3. Non-Binder configurations exhibit strong correlations with reduction clusters, though to a lesser degree than binders. *Identified SPACE2 configurations in example UMAP-correlation with their associated reduction clusters. Contains the number of CDRH3 described by each configuration and the percentage of the associated reduction cluster it represents. Ordered from high to low by number of CDRH3 in each configuration. See Supplementary Table 2 for binders*

Binder Reduction Clusters (Excluding Out-of-Bounds Clusters)				
Cluster Name	***# of CDRH3s***	***# of Configurations***	***Priority***	***Average Configuration RMSD***
Binder10	***1772***	***2***	***high***	***3.92***
Binder0	***1690***	***1***	***high***	***N/A***
Binder5	***970***	***2***	***high***	***2.64***
Binder4	***949***	***2***	***high***	***1.69***
Binder14	***318***	***1***	***med***	***N/A***
Binder2	***261***	***1***	***med***	***N/A***
Binder7	***204***	***1***	***med***	***N/A***
Binder16	***195***	***1***	***med***	***N/A***
Binder11	***193***	***1***	***med***	***N/A***
Binder22	***180***	***1***	***med***	***N/A***
Binder8	***152***	***1***	***med***	***N/A***
Binder26	***116***	***1***	***med***	***N/A***
Binder20	***109***	***1***	***med***	***N/A***
Binder31	***104***	***1***	***med***	***N/A***

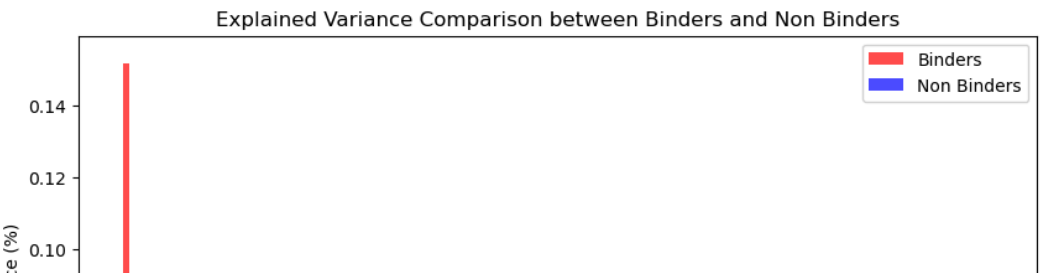
Binder27	*101*	*1*	*med*	*N/A*
Binder6	*97*	*1*	*med*	*N/A*
Binder9	*94*	*1*	*med*	*N/A*
Binder19	*93*	*1*	*med*	*N/A*
Binder23	*88*	*1*	*low*	*N/A*
Binder1	*86*	*1*	*low*	*N/A*
Binder3	*86*	*1*	*low*	*N/A*
Binder25	*78*	*1*	*low*	*N/A*
Binder13	*64*	*1*	*low*	*N/A*
Binder24	*59*	*1*	*low*	*N/A*
Binder17	*53*	*1*	*low*	*N/A*
Binder28	*50*	*1*	*low*	*N/A*
Binder29	*45*	*1*	*low*	*N/A*
Binder34	*44*	*1*	*low*	*N/A*
Binder15	*38*	*1*	*low*	*N/A*
Binder12	*36*	*1*	*low*	*N/A*
Binder18	*34*	*1*	*low*	*N/A*
Binder21	*27*	*1*	*low*	*N/A*
Binder33	*27*	*1*	*low*	*N/A*
Binder32	*23*	*1*	*low*	*N/A*
Binder30	*20*	*1*	*low*	*N/A*
Binder35	*16*	*1*	*low*	*N/A*

Supplementary table 4. Multiconfigurational binder reduction clusters exhibit a broad range of RMSD values. *Summary table of binder reduction clusters from example UMAP-correlation, including number of CDRH3s contained in each, number of configurations calculated by SPACE2, and RMSD values of those configurations, if applicable. Ordered from high to low by the number of CDRH3s in each cluster. See supplementary table 5 for non-binder reduction clusters and appendix_folder/data/umap for raw data*

Non Binder Reduction Clusters (Excluding Out-of-Bounds Clusters)

Cluster Name	# of CDRH3s	# of Configurations	Priority	Average Configuration RMSD
Non Binder6	1910	2	med	5.46
Non Binder4	1788	2	med	2.29
Non Binder3	1690	2	med	4.09
Non Binder13	1570	1	med	N/A
Non Binder8	1050	4	med	3.21
Non Binder10	839	1	med	N/A
Non Binder0	832	1	med	N/A
Non Binder2	747	1	med	N/A
Non Binder7	516	2	med	3.61
Non Binder5	513	1	med	N/A
Non Binder9	316	2	med	2.64
Non Binder-1	238	1	low	N/A
Non Binder11	231	1	low	N/A
Non Binder16	160	1	low	N/A
Non Binder17	119	1	low	N/A
Non Binder12	78	1	low	N/A
Non Binder18	29	1	low	N/A
Non Binder14	28	1	low	N/A
Non Binder15	24	1	low	N/A
Non Binder19	20	1	low	N/A

Supplementary Table 5. Multiconfigurational non-binder reduction clusters exhibit a broad range of RMSD values. *Summary table of binder reduction clusters from example UMAP-correlation, including number of CDRH3s contained in each, number of configurations calculated by SPACE2, and RMSD values of those configurations, if applicable. Ordered from high to low by the number of CDRH3s in each cluster. See supplementary table 4 for non-binder reduction clusters and appendix_folder/data/umap for raw data.*



Supplemental Figure 1.

Explained Variance for binders (red) and nonbinders (blue) remains broadly consistent past the first three components.

Supplemental Data

https://drive.google.com/drive/folders/1sAeUJA-Fh83lh-on8dzLE9q7MxyK0W5G?usp=sharing

Contains the following:

- ***Code**: All python scripts and Jupyter Notebooks used to generate the data*
- ***Data**: All the results used in this study and report. Subfolders and files are as follows:*
    - ***Logoplots**: All the logoplots generated throughout the course of this study, with subfolders for the full library and UMAP cluster analysis, with all color schemes utilized*
    - ***mapping_PCAdatatoPCAgraphs:** Folder containing PCA derived configurations mapped onto original PCA reduction graphs, in interactive plotly format*
    - ***mapping_PCAdatatoUMAPgraphs:** Folder containing UMAP-correlation derived configurations mapped onto original PCA reduction graphs, in interactive plotly format*
    - ***mapping_PCAdatatoUMAPgraphs:** Folder containing UMAP derived configurations mapped onto original UMAP reduction graphs, in interactive plotly format.*
    - ***PCA:** Contains data used in PCA data, including original coordinates of first 95 components, pairplots of first 20 principal components, the results of the SPACE2 analysis across all reductions, sequences with all PCA derived configurations, sequences with PCA derived configuration supercluster.*
    - ***UMAP:** Contains data used in PCA data, including original coordinates of first all reductions analyzed, the results of the SPACE2 analysis across all reductions, sequences with all UMAP derived configurations, sequences with UMAP derived configuration superclusters.*
