EVALUATING PREDICTIVE PATTERNS OF ANTIGEN SPECIFIC B CELLS BY SINGLE CELL TRANSCRIPTOME AND ANTIBODY REPERTOIRE SEQUENCING

Lena Erlach, Raphael Kuhn, Andreas Agrafiotis, Danielle Shlesinger,
Alexander Yermanos & Sai T. Reddy
Department of Biosystems Science and Engineering
ETH Zürich
Basel, Switzerland
{lena.erlach, sai.reddy}@bsse.ethz.ch

ABSTRACT

The field of antibody drug discovery relies substantially on extensive experimental screening of B cells from immunized animals. Machine learning (ML)-guided prediction of antigen-specific B cells offers the potential to accelerate antibody drug discovery, however this requires sufficient labeled training data. Addressing this challenge, our study focuses on antigen specificity prediction using a novel dataset of B cells with single-cell transcriptome and antibody repertoire sequencing. We identify key patterns in gene expression (GEX) indicative of antigen specificity and elucidate the sequence diversity distribution of antigen-specific antibody sequences in immune repertoire data. We evaluate linear (Logistic Regression), non-linear (Support Vector Classification) and ensemble-based (Random Forest, Gradient Boosting) models trained on different feature combinations of GEX and antibody sequence. Additionally, transfer learning approaches using features generated from ESM-2, a general protein language model (PLM), as well as from AntiBERTy, an antibody specific PLM, were evaluated as inputs to these models. Our findings reveal that GEX-based models demonstrate superior performance in specificity predictions with F1 scores up to 0.939 compared to antibody sequence-based models, highlighting the intricate nature of immune repertoire modeling. Contrary to our expectations, using PLM features did not enhance predictive accuracy. Our research contributes to the computational discovery of antibody therapeutics, offering insights into B cell biology and serving as dataset contribution to the development of ML approaches in this field.

1 INTRODUCTION

Therapeutic antibodies have become one of the most dominant drug modalities for the biopharmaceutical industry (Kaplon et al., 2023). Many antibody discovery campaigns rely on animal immunization and experimental screening of B cells (Kellermann & Green, 2002; Laustsen et al., 2021) (Figure 1A). A critical aspect of this process is the identification of B cells expressing antibodies that not only exhibit high specificity and affinity towards the target antigen, but also meet key biophysical criteria (developability), which are important for downstream drug development (Jain et al., 2017; Raybould et al., 2019; Jarasch et al., 2015). Utilizing animal immunization, while technologically straightforward, offers the advantage of producing antibodies that have undergone *in vivo* affinity maturation, yielding high affinity and *in vivo* stability (Laustsen et al., 2021). However, downstream engineering steps often require extensive wet lab validation of 1,000 - 10,000 antibodies, while still yielding only a small subset of therapeutic candidates (<100). (Zost et al., 2020; Chi et al., 2020; Brouwer et al., 2020). Recent technological advancements in single-cell sequencing, such as LIBRA-seq, have enhanced the efficiency of discovering antigen-specific antibodies from immune repertoires. LIBRA-seq is a B cell receptor sequencing approach that utilizes DNA-barcoded antigens to map antibody sequence to antigen specificity using next-generation sequencing (Shiakolas et al., 2022; Setliff et al., 2019). While potentially increasing throughput, these methods still require extensive experimental screening, which can again hinge on the availability of expressible antigens and the steric accessibility of the epitopes. The potential to use computational tools, such as ML to identify antigen-specific antibodies from immune repertoires has gained substantial interest due to their potential to considerably reduce wet lab experimentation costs and efforts (Akbar et al., 2022). Nevertheless, the success of these methods is still hampered by insufficient data of labeled (antigen-specific) antibody sequences (Greiff et al., 2020).

To address these challenges, we have generated a dataset of antigen-specific and non-specific B cells with single-cell transcriptome and antibody repertoire sequencing. We hypothesize that post-activation gene expression patterns in B cells, as well as convergent patterns in antibody sequences are indicative of antigen specificity(Young & Brink, 2021; Cyster & Allen, 2019). By leveraging training data of transcriptional B cell phenotypes and antibody sequences, we aim to improve the accuracy of antigen-specificity prediction models. Thus, in this study we evaluate ML models trained on this dataset that serve as benchmarks to identify learnable patterns in the transcriptome and antibody sequences associated with antigen specificity. While our primary goal is to enhance ML methods in antibody discovery, the dataset also offers valuable insights into B cell biology. Ultimately, by demonstrating the efficacy of PLM embeddings in predicting antigen specificity, we anticipate this work to contribute to the computational design of future antibody therapeutics, leveraging the generative capabilities of PLMs (Shuai et al., 2023; Wang et al., 2023).

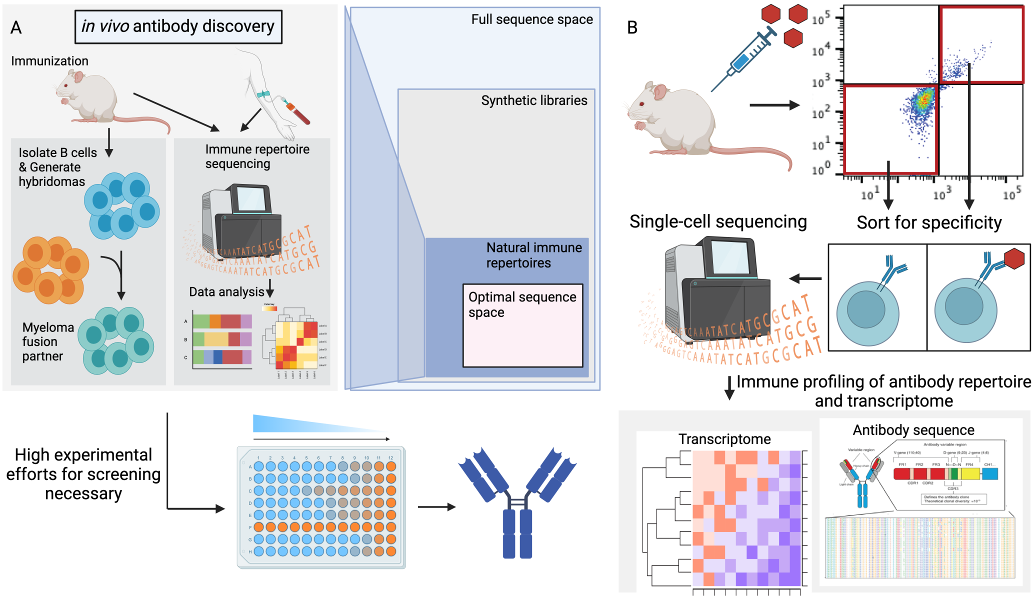
A
in vivo antibody discovery
Immunization
Isolate B cells & Generate hybridomas
Immune repertoire sequencing
Myeloma fusion partner
Data analysis
Full sequence space
Synthetic libraries
Natural immune repertoires
Optimal sequence space
High experimental efforts for screening necessary
B
Single-cell sequencing
Sort for specificity
Immune profiling of antibody repertoire and transcriptome
Transcriptome
Antibody sequence

Figure 1: A. *In vivo* antibody discovery technologies are reliant on extensive experimental screening, but yield antibodies with beneficial biophysical properties. B. Dataset Generation. Mice were immunized with antigens. Organs of the immunized animals were processed and single cells separated in antigen-specific and non-specific cell populations. Transcriptome and antibody genes of the single-cell populations were profiled by single-cell transcriptome and antibody repertoire sequencing.

2 Methods

2.1 Data generation

One of the major bottlenecks in the development and validation of computational methods for antibody discovery and engineering is the lack of sufficient labeled datasets. Therefore, we aimed to generate single-cell transcriptome and antibody sequence data from B cells of antigen-immunized mice. The experimental workflow is depicted in Figure 1B. We immunized two groups of mice, each with one of two protein antigens: ovalbumin (OVA) and the receptor binding domain (RBD) of

SARS-CoV-2. Following immunization, B cells were sorted by flow cytometry into antigen-specific and non-specific populations. Single-cell sequencing was performed on the sorted cell fractions, resulting in two primary datasets. Each dataset was characterized by three key features or labels: (1) a binary for antigen-specific and non-specific cells, (2) GEX, and (3) antibody sequences. Notably, sequencing for the OVA-immunized samples was conducted individually for each mouse, generating separate datasets for specific and non-specific cells (referred to as s1_OVA_spec and s2_OVA_nonspec for mouse 1, and s3_OVA_spec and s4_OVA_nonspec for mouse 2). For the RBD-immunized mice, samples from both individuals were pooled, resulting in a consolidated dataset for specific cells (labeled s1_RBD_spec) and another for non-specific cells (labeled s2_RBD_nonspec). Details on the performed experiments can be found in the appendix A.

2.2 Specificity prediction benchmarks

To investigate the presence of learnable patterns predictive for antigen specificity within our datasets, we conducted benchmark experiments using ML models that included linear (Logistic Regression, LogReg), non-linear (kernel Support Vector Machine Classifier, kSVC) and ensemble models (Random Forest, RF; Gradient Boosted Decision trees, GBoost (Ho, 1995; Friedman, 2002)). These binary classification models were trained on features derived from GEX and antibody sequences, within a framework of 5-fold cross-validation. Train-test splits were either sampled randomly in a stratified manner (random splits) or split based on antibody sequence similarity, thus preventing highly similar sequences in both training and test sets (similarity based splits). We refer to A.7 for details on train test splitting. In addition, we also leveraged transfer learning, a concept supported by its success in predicting protein function (Rao et al., 2019; Hie et al., 2020; Li et al., 2023). The rationale behind this approach stems from the hypothesis that features derived from PLMs that are trained on vast repositories of protein sequences could potentially enable our models to tap into a broader "protein universe" for improved prediction of protein function (Hie et al., 2023). Therefore, representations were generated from the last hidden layer states of two PLMs, ESM-2 and AntiBERTy, and used as input features for our ML models. ESM-2 and AntiBERTy are deep contextual language models containing information about biological properties of proteins and antibodies. ESM-2 was trained with 250 million protein sequences (Rives et al., 2021) and AntiBERTy is an antibody specific PLM, that was trained with 558 million natural antibody sequences (Ruffolo et al., 2021).

Feature generation: For specificity prediction, we generated diverse feature sets from GEX and antibody sequence data. GEX features are the scaled, log-transformed gene expression of the top 2000 most variable genes (GEX_2000_var) that are fed as input to the ML models. These features were prepared separately for the OVA, the RBD and an integrated OVA_RBD dataset (using Harmony single cell integration method (Korsunsky et al., 2019)). Using the integrated dataset for the specificity prediction of two distinct antigens follows the assumption that GEX patterns in activated B cells are similar across antigens. Antibody sequence features included k-mer frequencies, specifically 3-mer, for both heavy and light chains (VH_VL) and heavy chain only (VH). Additionally, PLM-derived features were obtained by mean-pooling the last hidden layer of ESM-2 (Rives et al., 2021) and AntiBERTy (Ruffolo et al., 2021) for both full-length VH and VL sequences or VH only. Furthermore, ESM-2 representations of complementary determining regions (CDR) (ESM-CDRextract) only were evaluated. This approach aimed to assess whether focusing on the CDRs, which are crucial for antigen binding, would yield sufficient or more effective embeddings for our specificity prediction models. To assess the synergy between GEX and antibody sequence data, GEX_2000_var and 3-mer features were concatenated and used as inputs to the ML models, exploring potential enhancement in predictive performance.

3 Results

3.1 Dataset

In response to immune challenges such as infections or immunizations, antigen-specific B cells undergo activation, followed by expansion and affinity maturation, a highly regulated, cellular process (Young & Brink, 2021). We propose that these antigen-specific B cells exhibit unique gene expression profiles that, once identified, could serve as a basis for predicting antigen specificity. Our single-cell sequencing experiments yielded a dataset encompassing 10,881 (OVA) and 7,677 (RBD) single-cell transcriptomes, alongside 8,226 (OVA) and 6,846 (RBD) paired heavy and light chain antibody sequences from individual cells. For visualization of our integrated dataset, we employed uniform manifold approximation and projection (UMAP) and labelled cells as OVA or RBD specific and non-specific. We observed limited clustering of antigen-specific B cells within the UMAP landscapes, indicating the presence of OVA and RBD-specific B cells across all clusters with a slightly lesser frequency towards the right side (Figure 2A). Additional analysis was performed and summarized in the Appendix A.5.

An expanding B cell and its descendants and the antibodies they express are assigned to a clonotype in immune repertoire analysis. We delineated clonotypes based on 10x Genomics Cell Ranger software (v5.0.0), which groups B cell clonotypes using enclone. The process of expansion leads to a skewed distribution characterized by large proportions of duplicated or very similar antibody sequences in immune repertoire datasets. Visualization in Figure 2B highlights the expansion of antigen-specific samples exhibiting a higher number of expanded clones, as evidenced by fewer grey fractions, thereby occupying a larger proportion of the dataset.

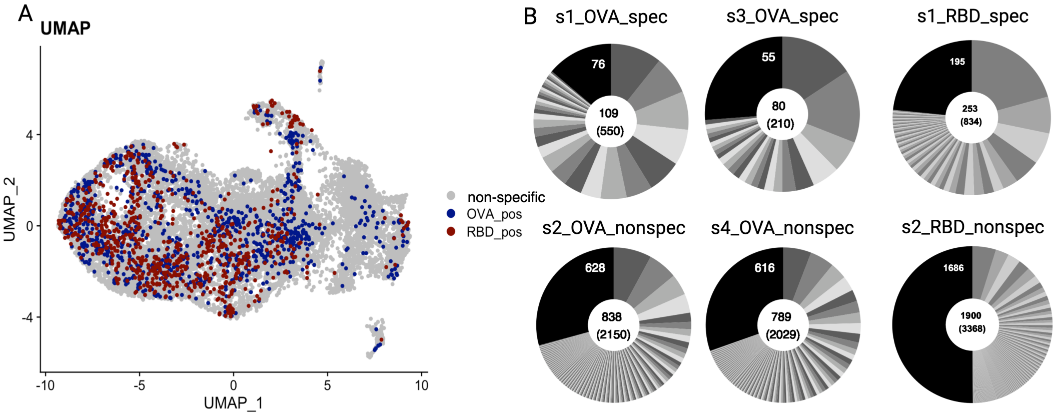

Figure 2: A. UMAP plot of single cells coloured by OVA, RBD specificity (blue, red); non-specific B cells depicted in grey. B. Pie charts indicating the fraction of B cells per clone and sample captured by single-cell sequencing. Numbers in the center indicate the total number of clones (Top) and cells (Bottom). Fraction of unexpanded clones consisting of one cell are shown in black and the fractions in varying shades of grey depict clones consisting of >1 cells.

3.2 Specificity prediction based on antibody sequences

Given the aim of predicting the positive class (antigen specific) as correctly as possible in this imbalanced dataset, F1 score was adopted as the primary metric for evaluating model performance (Figure 3A). For similarity based splits, the best-performing models were identified as LogReg trained on ESM-2 features for the OVA dataset (F1=0.570) and AntiBERTy features for the RBD dataset (F1=0.408), as detailed in Table 3. As expected, when the train-test datasets were split randomly, a significant improvement in predictive performance was observed (Figure 6A). Hereby, the top-performing model was RF trained on 3-mer features achieving F1 scores of 0.795 for OVA and 0.690 for RBD. When the cutoff of the similarity based splits was chosen more stringently, a decrease in performance was observed (See Table 8). These results highlight the impact of dataset splitting strategies on model efficacy in the context of antibody specificity predictions. Interestingly, transfer learning approaches using PLM representations of ESM-2 and AntiBERTy as input features did not result in a significant increase in F1 scores compared to 3-mer features (See Section A.10, Figure 6A). However, LogReg trained on ESM-2 and AntiBERTy features of the OVA and RBD dataset, respectively, exhibited the best performance in similarity based splits, suggesting enhanced generalization capabilities to more distantly related sequences. Unexpectedly, ML models trained with features from AntiBERTy, an antibody specific PLM, did not outperform models trained with ESM-2 features (Figure 6A). Furthermore, when CDR representations were extracted from the full-length PLM representation and used as features, we observed either equal or slightly improved performance for the specificity prediction task. This finding indicates that the contextual information captured in the CDRextract features, as opposed to the full-length sequence ESM-2 features, appears sufficient for specificity prediction. Even though more detailed evaluations would be required, these insights could imply that it is sufficient to generate PLM features of CDRs only. Shorter sequences lead to more efficient generation of PLM features, a process typically demanding significant computational resources.

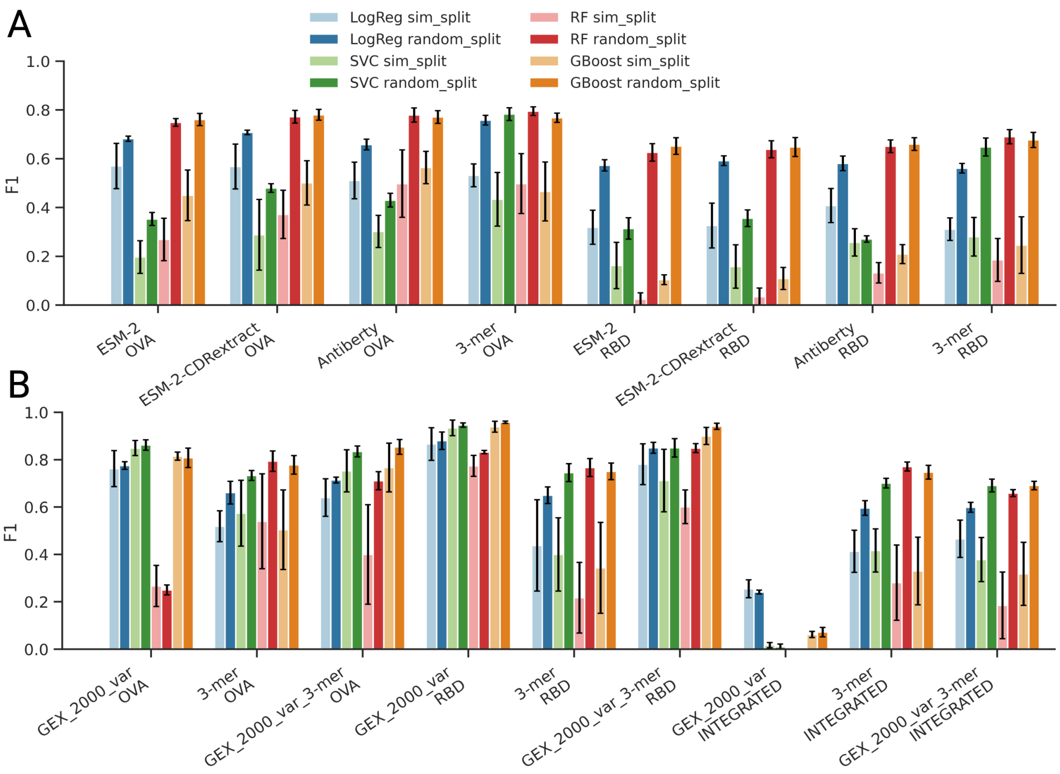

Figure 3: F1 score of model evaluations based on A) antibody sequence features and B) GEX features.

3.3 Specificity prediction based on gene expression

In this analysis, we focused on cells where both GEX and antibody sequence data was available, to ensure a consistent basis for evaluation of the specificity predictions. Additionally, we included sequence-only feature (3-mer) predictions for these datasets to provide a comparative perspective. The F1 scores of the model predictions (Figure 3B and Table 9) reveal notable predictive performance of GEX features in antigen specificity prediction. This finding reaffirms the results in the analysis of differentially expressed (DE) genes (Figure 5). Specifically, using kSVC with solely GEX features yielded remarkable F1 scores of up to 0.849 for OVA and 0.939 for RBD. GEX features for the RBD dataset yield statistically significant improvements compared to sequence based features. Interestingly, when GEX and antibody sequence features were combined, this did not enhance performance for the OVA and RBD datasets. This outcome suggests superiority of GEX features over antibody sequence features in the context of antigen specificity prediction, particularly in the RBD dataset (Figure 3B, Figure 7B). Further, our analysis revealed only a minor, but not significant decrease in F1 scores when employing similarity based splits compared to random splits in GEX predictions. This observation suggests that splitting train and test sets based on sequence similarity, or lackthereof, does not substantially affect the predictive power of GEX features (Figure 7A).

In contrast, when predictions on the integrated dataset were evaluated, the effectiveness of GEX features to predict antigen specificity was significantly reduced. Further analysis of the coefficient values learned by the LogReg models trained on separate OVA and RBD datasets showed only minor correlation and overlap (Figure 8, 9 and see A.12 for details). Even though these results imply divergent antigen-specific gene signatures that are learned by the LogReg models to predict antigen specificity, more detailed analysis would be required to confirm this hypothesis. Nevertheless, combined GEX and antibody features did exhibit a capacity to recover some predictive performance, achieving an F1 score of 0.466. While this score marks an improvement compared to GEX-only models, it falls notably short of the F1 scores obtained for the model evaluations when trained on the separate OVA and RBD datasets, especially compared within similarity based splits. Eventually, a control experiment was conducted using the top 10 upregulated and downregulated differentially expressed (DE) genes to evaluate the necessity of ML in predicting antigen specificity from GEX features with the integrated OVA and RBD dataset (See Appendix A.13 for details). These results (Table 9) show slightly improved F1 scores compared to model predictions based on GEX features for the OVA_RBD dataset, but cannot outperform combined GEX_3-mer features.

4 Discussion

As proof-of-concept, we demonstrate that antigen specificity predictions of B cells based on single cell GEX data are effective within an antigen cohort, yet susceptible when applied to integrated datasets of two antigen cohorts. A hypothesis for the low performance of the ML models using integrated datasets could be a divergence in antigen-specific gene expression, challenging our assumption of a 'universal' antigen-specific response. These findings are supported by our GEX analysis, which shows separation on the PCA plot based on specificity as well as antigen (Figure 5A). The analysis of the UMAP plot, however, is in contrast to these results as it does not depict any distinct clustering or separation of B cells activated by two different antigens (Figure 2A). Given that UMAP radically reduces the dimensionality of GEX, it may not fully capture complex GEX pattern. The superior performance of nonlinear models such as kSVC and GBoost when using GEX features, highlights the complexity and intricacy of antigen-specific GEX signatures. These diverging results in model performance also suggest the presence of random batch effects, which could be investigated by generating additional, independent batches of single cell sequencing datasets. Such investigations require carefully designed experiments incorporating both technical and biological replicates. This remains technically challenging, however, due to limitations in the throughput of animal experiments and the number of cells that can be sequenced in a single experiment. Further experiments are also required to experimentally validate the results of the model predictions. To our current knowledge, no study has thoroughly explored differences in antigen specific GEX patterns of primary B cells.

The results of the antibody sequence-based predictions clearly showed a significant impact of the similarity split threshold on on the overall predictive performance, underscoring the complexities inherent in modeling immune repertoire datasets that are shaped by natural immune responses, such as clonal expansion. Interestingly, transfer learning by using PLM representations, such as those from ESM-2, did not enhance model performance. Similarly, when employing ML models with representations from AntiBERTy, a PLM specifically trained with antibody sequences, no improvement in model performance was observed. These results stand in contrast to other studies that have documented improvements of downstream protein function prediction using PLM representations (Rao et al., 2019; Hie et al., 2020; Li et al., 2023). Nonetheless, the observed predictive capacity of PLM features for antigen specificity opens avenues for further investigation into how PLMs could be utilized for the design of novel antibody sequences. This could entail sampling from the PLM embedding space or leveraging PLM pseudo log-likelihoods to guide antibody design (Hie et al., 2023). Moreover, our findings indicate that combining GEX and antibody sequence features does not improve model performance for antigen specificity prediction. More sophisticated methods than simple feature concatenation, however, could further enhance model performance. One such approach could utilize features from a latent space that integrates antibody sequence and GEX data, such as the Benisse model (Zhang et al., 2022). Taken together this study provides critical insights into the challenges and potential directions for advancing ML applications in antibody discovery from immune repertoires.

References

Rahmad Akbar, Habib Bashour, Puneet Rawat, Philippe A. Robert, Eva Smorodina, Tudor Stefan Cotet, Karine Flem-Karlsen, Robert Frank, Brij Bhushan Mehta, Mai Ha Vu, Talip Zengin, Jose

Gutierrez-Marcos, Fridtjof Lund-Johansen, Jan Terje Andersen, and Victor Greiff. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs*, 14(1), dec 2022. ISSN 19420870. doi: 10.1080/19420862.2021.2008790. URL https://www.tandfonline.com/doi/abs/10.1080/19420862.2021.2008790.

Dmitriy A. Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z. Mamedov, Ekaterina V. Putintseva, and Dmitriy M. Chudakov. MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods 2015 12:5*, 12(5):380–381, apr 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3364. URL https://www.nature.com/articles/nmeth.3364.

Philip J.M. Brouwer, Tom G. Caniels, Karlijn van der Straten, Jonne L. Snitselaar, Yoann Aldon, Sandhya Bangaru, Jonathan L. Torres, Nisreen M.A. Okba, Mathieu Claireaux, Gius Kerster, Arthur E.H. Bentlage, Marlies M. van Haaren, Denise Guerra, Judith A. Burger, Edith E. Schermer, Kirsten D. Verheul, Niels van der Velde, Alex van der Kooi, Jelle van Schooten, Mariëlle J. van Breemen, Tom P.L. Bijl, Kwinten Sliepen, Aafke Aartse, Ronald Derking, Ilja Bontjer, Neeltje A. Kootstra, W. Joost Wiersinga, Gestur Vidarsson, Bart L. Haagmans, Andrew B. Ward, Godelieve J. de Bree, Rogier W. Sanders, and Marit J. van Gils. Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science*, 369 (6504):643–650, aug 2020. ISSN 10959203. doi: 10.1126/SCIENCE.ABC5902/SUPPL_FILE/ABC5902-BROUWER-SM.PDF. URL https://www.science.org/doi/10.1126/science.abc5902.

Xiangyang Chi, Renhong Yan, Jun Zhang, Guanying Zhang, Yuanyuan Zhang, Meng Hao, Zhe Zhang, Pengfei Fan, Yunzhu Dong, Yilong Yang, Zhengshan Chen, Yingying Guo, Jinlong Zhang, Yaning Li, Xiaohong Song, Yi Chen, Lu Xia, Ling Fu, Lihua Hou, Junjie Xu, Changming Yu, Jianmin Li, Qiang Zhou, and Wei Chen. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*, 369(6504):650–655, aug 2020. ISSN 10959203. doi: 10.1126/SCIENCE.ABC6952/SUPPL_FILE/ABC6952_S1.MP4. URL https://www.science.org/doi/10.1126/science.abc6952.

Tudor Stefan Cotet, Andreas Agrafiotis, Victor Kreiner, Raphael Kuhn, Danielle Shlesinger, Marcos Manero-Carranza, Keywan Khodaverdi, Evgenios Kladis, Aurora Desideri Perea, Dylan Maassen-Veeters, Wiona Glänzer, Solène Massery, Lorenzo Guerci, Kai Lin Hong, Jiami Han, Kostas Stiklioraitis, Vittoria Martinolli D'Arcy, Raphael Dizerens, Samuel Kilchenmann, Lucas Stalder, Leon Nissen, Basil Vogelsanger, Stine Anzböck, Daria Laslo, Sophie Bakker, Melinda Kondorosy, Marco Venerito, Alejandro Sanz García, Isabelle Feller, Annette Oxenius, Sai T. Reddy, and Alexander Yermanos. ePlatypus: an ecosystem for computational analysis of immunogenomics data. *Bioinformatics*, 39(9), sep 2023. ISSN 13674811. doi: 10.1093/BIOINFORMATICS/BTAD553. URL https://dx.doi.org/10.1093/bioinformatics/btad553.

Jason G. Cyster and Christopher D.C. Allen. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell*, 177(3):524–540, apr 2019. ISSN 0092-8674. doi: 10.1016/J.CELL.2019.03.016.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics Data Analysis*, 38(4): 367–378, feb 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2.

Victor Greiff, Gur Yaari, and Lindsay G. Cowell. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, 24:109–119, dec 2020. ISSN 2452-3100. doi: 10.1016/J.COISB.2020.10.010.

Brian Hie, Bryan D. Bryson, and Bonnie Berger. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell systems*, 11(5):461–477.e9, nov 2020. ISSN 2405-4720. doi: 10.1016/J.CELS.2020.09.007. URL https://pubmed.ncbi.nlm.nih.gov/33065027/.

Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U.J. Bruun, Payton A. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak, and Peter S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology 2023*, pp. 1–9, apr 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01763-2. URL https://www.nature.com/articles/s41587-023-01763-2.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, 1995.

Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H. Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T. Gray, Eric M. Krauland, Yingda Xu, Maximiliano Vásquez, and K. Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 114(5):944–949, jan 2017. ISSN 10916490. doi: 10.1073/PNAS.1616408114/SUPPL_FILE/PNAS.1616408114.SD03.XLSX. URL https://www.pnas.org/doi/abs/10.1073/pnas.1616408114.

Alexander Jarasch, Hans Koll, Joerg T. Regula, Martin Bader, Apollon Papadimitriou, and Hubert Kettenberger. Developability Assessment During the Selection of Novel Therapeutic Antibodies. *Journal of Pharmaceutical Sciences*, 104 (6):1885–1898, jun 2015. ISSN 0022-3549. doi: 10.1002/JPS.24430. URL http://jpharmsci.org/article/S0022354915300848/fulltexthttp://jpharmsci.org/article/S0022354915300848/abstracthttps://jpharmsci.org/article/S0022-3549(15)30084-8/abstract.

Hélène Kaplon, Silvia Crescioli, Alicia Chenoweth, Jyothsna Visweswaraiah, and Janice M. Reichert. Antibodies to watch in 2023. *mAbs*, 15(1), dec 2023. ISSN 19420870. doi: 10.1080/19420862.2022.2153410. URL https://www.tandfonline.com/doi/abs/10.1080/19420862.2022.2153410.

Sirid Aimée Kellermann and Larry L. Green. Antibody discovery: the use of transgenic mice to generate human monoclonal antibodies for therapeutics. *Current Opinion in Biotechnology*, 13 (6):593–597, dec 2002. ISSN 0958-1669. doi: 10.1016/S0958-1669(02)00354-3.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods 2019 16:12*, 16(12): 1289–1296, nov 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0. URL https://www.nature.com/articles/s41592-019-0619-0.

Andreas H. Laustsen, Victor Greiff, Aneesh Karatt-Vellatt, Serge Muyldermans, and Timothy P. Jenkins. Animal Immunization, in Vitro Display Technologies, and Machine Learning for Antibody Discovery. *Trends in Biotechnology*, 39(12):1263–1273, dec 2021. ISSN 18793096. doi: 10.1016/j.tibtech.2021.03.003. URL http://www.cell.com/article/S0167779921000615/fulltexthttp://www.cell.com/article/S0167779921000615/abstracthttps://www.cell.com/trends/biotechnology/abstract/S0167-7799(21)00061-5.

Lin Li, Esther Gupta, John Spaeth, Leslie Shing, Rafael Jaimes, Emily Engelhart, Randolph Lopez, Rajmonda S Caceres, Tristan Bepler, and Matthew E Walsh. Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nature Communications*, 14(1):3454, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39022-2. URL https://doi.org/10.1038/s41467-023-39022-2.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12): 1–21, dec 2014. ISSN 1474760X. doi: 10.1186/S13059-014-0550-8/FIGURES/9. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

Daniel Neumeier, Alexander Yermanos, Andreas Agrafiotis, Lucia Csepregi, Tasnia Chowdhury, Roy A. Ehling, Raphael Kuhn, Tudor Stefan Cotet, Raphaël Brisset-Di Roberto, Mariangela Di Tacchio, Renan Antonialli, Dale Starkie, Daniel J. Lightwood, Annette Oxenius, and Sai T. Reddy. Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *Proceedings of the National Academy of Sciences of the United States of America*, 119(18), may 2022. ISSN 10916490. doi: 10.1073/PNAS.2113766119/-/DCSUPPLEMENTAL. URL https://pubmed.ncbi.nlm.nih.gov/35486691/.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32:9689, 2019. ISSN 10495258. URL /pmc/articles/PMC7774645//pmc/articles/PMC7774645/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7774645/.

Matthew I.J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4025–4030, 2019. ISSN 10916490. doi: 10.1073/PNAS.1810576116/-/DCSUPPLEMENTAL. URL www.pnas.org/cgi/doi/10.1073/pnas.1810576116.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118:e2016239118, 4 2021. ISSN 10916490. doi: 10.1073/PNAS.2016239118/SUPPL_FILE/PNAS.2016239118.SAPP.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. dec 2021. URL https://arxiv.org/abs/2112.07782v1.

Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology 2015 33:5*, 33(5):495–502, apr 2015. ISSN 1546-1696. doi: 10.1038/nbt.3192. URL https://www.nature.com/articles/nbt.3192.

Ian Setliff, Andrea R. Shiakolas, Kelsey A. Pilewski, Amyn A. Murji, Rutendo E. Mapengo, Katarzyna Janowska, Simone Richardson, Charissa Oosthuysen, Nagarajan Raju, Larance Ronsard, Masaru Kanekiyo, Juliana S. Qin, Kevin J. Kramer, Allison R. Greenplate, Wyatt J. McDonnell, Barney S. Graham, Mark Connors, Daniel Lingwood, Priyamvada Acharya, Lynn Morris, and Ivelin S. Georgiev. High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell*, 179(7):1636–1646.e15, dec 2019. ISSN 1097-4172. doi: 10.1016/J.CELL.2019.11.003. URL https://pubmed.ncbi.nlm.nih.gov/31787378/.

Andrea R. Shiakolas, Kevin J. Kramer, Nicole V. Johnson, Steven C. Wall, Naveenchandra Suryadevara, Daniel Wrapp, Sivakumar Periasamy, Kelsey A. Pilewski, Nagarajan Raju, Rachel Nargi, Rachel E. Sutton, Lauren M. Walker, Ian Setliff, James E. Crowe, Alexander Bukreyev, Robert H. Carnahan, Jason S. McLellan, and Ivelin S. Georgiev. Efficient discovery of SARS-CoV-2-neutralizing antibodies via B cell receptor sequencing and ligand blocking. *Nature Biotechnology 2022 40:8*, 40(8):1270–1275, mar 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01232-2. URL https://www.nature.com/articles/s41587-022-01232-2.

Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. Iglm: Infilling language modeling for antibody sequence design. *Cell Systems*, 14:979–989.e4, 11 2023. ISSN 2405-4712. doi: 10.1016/J.CELS.2023.10.001.

Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J.E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, Michael A. Skinnider, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications 2021 12:1*, 12(1):1–15, sep 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25960-2. URL https://www.nature.com/articles/s41467-021-25960-2.

Danqing Wang, Fei Ye, and Hao Zhou. On pre-trained language models for antibody. 1 2023. URL https://arxiv.org/abs/2301.12112v2.

Clara Young and Robert Brink. The unique biology of germinal center b cells. *Immunity*, 54:1652–1664, 8 2021. ISSN 1074-7613. doi: 10.1016/J.IMMUNI.2021.07.015.

Ze Zhang, Woo Yong Chang, Kaiwen Wang, Yuqiu Yang, Xinlei Wang, Chen Yao, Tuoqi Wu, Li Wang, and Tao Wang. Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse. *Nature Machine Intelligence 2022 4:6*, 4(6):596–604, jun 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00492-6. URL `https://www.nature.com/articles/s42256-022-00492-6`.

Seth J. Zost, Pavlo Gilchuk, Rita E. Chen, James Brett Case, Joseph X. Reidy, Andrew Trivette, Rachel S. Nargi, Rachel E. Sutton, Naveenchandra Suryadevara, Elaine C. Chen, Elad Binshtein, Swathi Shrihari, Mario Ostrowski, Helen Y. Chu, Jonathan E. Didier, Keith W. MacRenaris, Taylor Jones, Samuel Day, Luke Myers, F. Eun-Hyung Lee, Doan C. Nguyen, Ignacio Sanz, David R. Martinez, Paul W. Rothlauf, Louis Marie Bloyet, Sean P.J. Whelan, Ralph S. Baric, Larissa B. Thackray, Michael S. Diamond, Robert H. Carnahan, and James E. Crowe. Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. *Nature Medicine 2020 26:9*, 26(9):1422–1427, jul 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0998-x. URL `https://www.nature.com/articles/s41591-020-0998-x`.

A Appendix

A.1 Mouse experiments

For protein immunizations, 6-week-old female C57BL/6 mice were repeatedly immunized 3 times in 9 days intervals s.c. into the flank with 100 μg of OVA (Sigma, A5503) or 10 μg RBD (Sino Biological, 40592-V08B) with 20 μg MPLA (Sigma, L6895) adjuvant and euthanized 11 days (OVA) and 16 days (RBD) post immunization. Euthanization was conducted by CO_2 asphyxiation and cervical dislocation and axillary and inguinal lymph nodes on the side of immunization were collected.

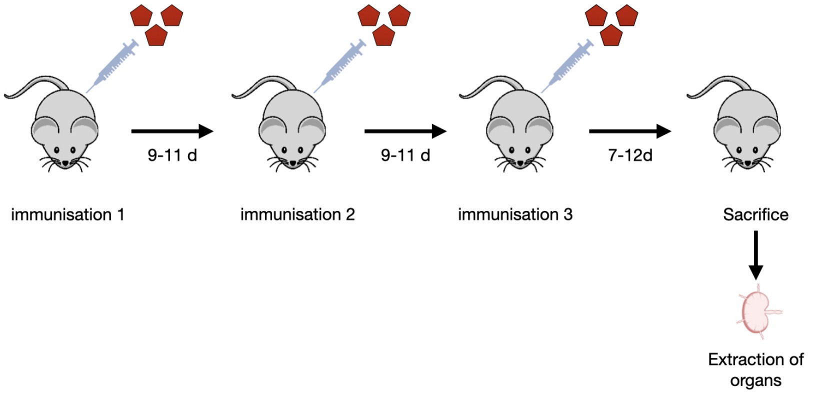

Figure 4: Immunization scheme of the mouse experiments

A.2 Fluourescence activated cell sorting

To prepare single cell suspensions, lymph nodes were mashed through a 70μm cell strainer. Single-cell suspensions were stained with fluourescently labelled with anti-CD19, anti-IGM, anti-IGD and fluourescently labelled target antigen (OVA, RBD). Cells were sorted on a FACSAria 3 (BD Biosciences) as activated B cells (CD19+, IGM-, IGD-) that are specific (OVA+/RBD+) or non-specific (OVA-/RBD-).

A.3 Single-cell sequencing and bioinformatic preprocessing

Single-cell immune repertoire and transcriptome sequencing was performed using the 10x Genomics Chromium Single-Cell V(D)J Reagents Kit (CG000166 Rev A) as previously described (Neumeier et al., 2022). In brief, single cells for all samples were processed using the 10x Genomics protocol and kits (User Guide, CG000207). After preparation of the sequencing sample, the GEX and VDJ library were pooled before sequencing on the Illumina NovaSeq S1 using a concentration of 1.8 pM with 5% PhiX. Paired-end sequencing files for GEX and VDJ libraries were aligned to the murine reference genome (mm10) and V(D)J germlines (GRCm38) using 10x Genomics cellranger (v5.0.0) software. Cell ranger output was further preprocessed and filtered using the R package Platypus (v3.2.1, (Cotet et al., 2023)), which uses the transcriptome analysis workflow of the R package Seurat (Satija et al., 2015). Only those cells containing less than 20% of mitochondrial reads were retained in the analysis. In addition, cells with barcodes that were not unique across samples of the same mouse were filtered. Genes involved in the adaptive immune receptor (e.g., IGH, IGK,...), were removed from the count matrix to prevent clonal relationships from influencing transcriptional phenotypes. Gene expression was normalized for each cell by the total expression, multiplied by a scale factor of 10,000, and log-transformed. 2000 variable features were selected using the “vst” selection method and used as input to principal component analysis (PCA, with 10 principle components). Dataset integration was performed by the single-cell integration package “harmony”, which was run on the sparse matrix of library size normalized expression counts. The package scales expression data, runs PCA, and the Harmony integration algorithm (Korsunsky et al., 2019). Graph-based clustering using the Louvain modularity optimization and hierarchical clustering was performed using the functions FindNeighbors and FindClusters in Seurat using the ten dimensions of the PCA and harmony output and a cluster resolution of 0.5. UMAP was performed equally. The antibody genes were realigend using MiXCR software v4.2.0 (Bolotin et al., 2015) for proper annotation according to AIRR guidelines.

A.4 Number of cells sequenced (GEX and antibody sequence)

In total we could successfully align 15,072 antibody heavy and light chain sequences of single cells in our dataset (Table 1). For the specificity prediction task we filtered for duplicated sequences, which are commonly observed in immune repertoire datasets because of expanded clones. In addition, we removed sequences that were seen in the specific, as well as non-specific samples. These sequences might have incorrect labels due do experimental measurement noise, and or exhibit low binding affinity.

In Table 2 the numbers of cells in our GEX dataset are summarized. After standard QC filtering and preprocessing we were able to successfully analyze the transcriptome of almost 18,558 cells in total. To reduce noise and biological variation in the data, filtering for T-cells and plasma cells (based on gene expression markers) was performed. In addition, we removed cells with identical CDR3H and CDR3Ls that were present in the specific and non-specific sample. For the specificity prediction evaluations based on GEX features, only cells were retained that had both, GEX and antibody sequence, profiled to ensure comparability across prediction tasks. Eventually 3,622 and 3,593 cells were retained in the OVA and RBD dataset, respectively.

Table 1: Antibody repertoire dataset size as number of antibody sequences

Dataset	specific	non-specific	Total
Before filtering			
OVA	823	7403	8226
RBD	1122	5724	6846
After filtering			
OVA	550	3072	3622
RBD	642	2951	3593

Table 2: GEX dataset size as number of cells's transcriptome profiled

Dataset	specific	non-specific	Total
Before filtering			
OVA	1071	9810	10881
RBD	1281	6296	7677
After filtering			
OVA	728	7912	8640
RBD	966	4874	5840
Cells with GEX and antibody sequence			
OVA	371	2937	3308
RBD	510	2051	2561
OVA_RBD_integrated	885	5029	5914

A.5 Differential gene expression analysis

To further evaluate our hypothesis that antigen-specific cells can be distinguished by their gene expression profiles, we conducted differential gene expression analysis (DGEA). Initially, pseudobulking was performed — a method demonstrated to enhance the robustness of detecting differentially expressed (DE) genes (Squair et al., 2021). Hence, gene expression counts were summed per sample followed by normalization and log-transformation. PCA revealed a segregation of specific versus non-specific samples along PC2, while PC1 differentiated between OVA and RBD samples Figure 2A, suggesting diverging signals across antigens or the presence of potential batch effects. We then applied DESeq2 (Love et al., 2014) to identify genes that were significantly upregulated or downregulated, identifying 859 genes with p-values less than 0.05. Notably, genes associated with cell cycle progression and DNA replication were upregulated, whereas those involved in lymphocyte activation and cytokine production regulation were downregulated (Figure 2B).

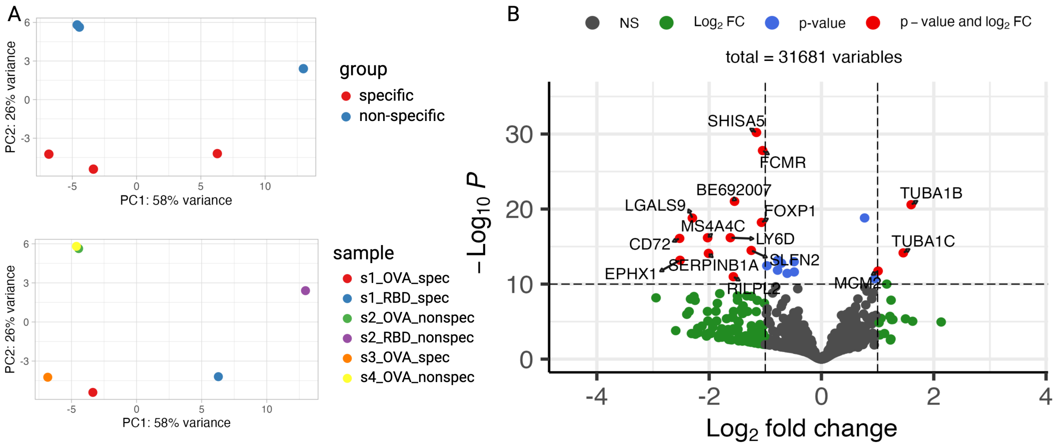

Figure 5: A. PCA of gene expression after pseudobulking mRNA transcripts per sample. B. Volcano plot of DE genes (specific vs. nonspecific) after pseudobulking mRNA transcripts per sample.

A.6 Details on feature generation

k-mer features: As antibody sequence features, we calculated k-mer frequecies, which were generated by identifying and counting all possible subsequences of length 3 within each sequence (3-mers). These counts were normalized by dividing by the total number of k-mers in each sequence, resulting in a frequency distribution for each k-mer. All unique k-mers from the dataset were aggregated and each sequence was represented as a vector, with elements corresponding to the frequencies of these k-mers. As the feature dimensions strongly increase with increasing k, 3-mers were chosen for our model evaluation. 3-mers represents a good trade-off between dimensionality of feature space and still capturing a more complex sequence representation compared to considering single amino acid frequencies.

PLM features: PLM representations were generated using the ESM-2 protein language model and the AntiBERTy language model. ESM-2 (esm2_t33_650M_UR50D) is a variant of the ESM model with 33 layers and 650 million parameters (Rives et al., 2021). This model was trained on the UR50/D 2021_04 dataset, as detailed in the ESM GitHub documentation: https://github.com/facebookresearch/esm?tab=readme-ov-file#available. AntiBERTy is an antibody-specific transformer language model, which was pre-trained on 558 million natural antibody sequences (Ruffolo et al., 2021). The documentation of AntiBERTy can be found in the following GitHub repository: https://github.com/jeffreyruffolo/AntiBERTy. The antibody protein sequences of both the heavy and light chains were tokenized and passed through the models. Thus, a PLM representation for each antibody sequence was generated by taking the last hidden layer's representation for each amino acid. The resulting variable length embedding, a sequence length-dependent vector of 1280 dimensions (seq_len x 1280) for ESM-1b and 512 dimensions for AntiBERTy (seq_len x 512) was then mean-pooled to obtain fixed-length embeddings for each antibody sequence (n_seqs x 1 x 1280). Specifically, for the heavy chain (VH) features, we used only the heavy chain embeddings, while for the VH_VL features, we concatenated and mean-pooled the representations from both the heavy and light chains.

ESM-CDRextract embeddings were generated by initially creating full-length representations and then selectively extracting tokens corresponding to the CDRs (CDRH1-3 and CDRL1-3). The extracted representations were mean-pooled to generate fixed-length feature vectors for both VH and VH_VL representations.

All of the features were preprocessed independently by removing the mean and scaling to unit variance before feeding into the ML models.

A.7 DETAILS ON TRAIN TEST SPLITTING

For the evaluation of our models, we implemented two distinct methods for generating train-test splits: random and sequence similarity-based splits. Random train test splits were generated using StratifiedShuffleSplit of the sklearn library. This function returns stratified randomized folds, maintaining the percentage of samples for both classes. This method ensures that each fold is a representative subset of the overall dataset.

To avoid data leakage by ensuring that highly similar sequences were not present in training and testing sets, we also evaluated our ML models on sequence similarity based train test splits. Therefore, we calculated normalized edit distances between each sequence and all the others in the dataset. Using the resulting distance matrix, hierarchical clustering (single linkage) was performed using the linkage function of the SciPy package. Clusters were then determined based on the linkage matrix, with the threshold parameter 't' resembling the edit distance cutoff that controls cluster assignment. This parameter acts as a cutoff, dictating the maximum allowed normalized edit distance for assigning sequences to a cluster. Based on the established clusters, we used the StratifiedGroupKFold function from sklearn to split the data into training and testing sets. The function attempts to return stratified folds of non-overlapping clusters. The folds are made by preserving the percentage of samples for each class.

A.8 IMPLEMENTATION OF ML MODELS

LogReg, kSVC, RF and GBoost models were all implemented using the sklearn library. The code is available on GitHub: https://github.com/LSSI-ETH/sc_AbSpecificity_pred.git. These models were assessed using nested cross validation for hyperparameter tuning and model evaluations. Parameter tuning was conducted with options detailed below in the parameter grids using RandomizedSearchCV also from the sklearn library, maximizing the metric recall. Each model's performance was evaluated using metrics like precision, recall, F1 score and AUC_ROC.

Parameter grid for LogReg: 'penalty': ['l2', None], 'class_weight': ['balanced', None], 'C': [10, 1, 0.1, 0.01, 0.001], 'max_iter': [500, 1000, 2000].

Parameter grid for kSVC: ’C’: [1, 0.1, 0.01, 0.001], ’kernel’: [’poly’, ’rbf’, ’sigmoid’], ’degree’: [2, 3, 4], ’gamma’: [’scale’, ’auto’].

Parameter grid for RF: ’n_estimators’: [100, 200, 300], ’max_depth’: [None, 5, 10], ’min_samples_split’: [2, 5, 10], ’min_samples_leaf’: [1, 2, 4], ’max_features’: [’auto’, ’sqrt’, ’log2’].

Parameter grid for GBoost: ’n_estimators’: [100, 200, 300], ’learning_rate’: [0.05, 0.1, 0.2], ’max_depth’: [3, 4, 5], ’min_samples_split’: [2, 5, 10], ’min_samples_leaf’: [1, 2, 4], ’max_features’: [’auto’, ’sqrt’, ’log2’, None], ’subsample’: [0.8, 0.9, 1.0].

A.9 Specificity predictions - Main tables

See Table 3 for a detailed summary of evaluation metrics: F1, precision (prec) and recall, of predictions with antibody sequence feature (heavy and light chain, VH_VL) with sequence similarity splits (distance threshold 0.05). Equally, Table 4 lists the detailed metrics of the ML models based on GEX features for specificity predictions (distance threshold 0.05).

Table 3: Specificity prediction results based on antibody sequence features - distance threshold for similarity based splits: 0.05

Dataset	**Similarity based split**				**random_split**			
OVA_VH_VL	F1	precision	recall	roc auc	F1	precision	recall	roc auc
LogReg_ESM	**0.570**	0.515	0.661	0.781	0.682	0.596	0.8	0.851
kSVC_ESM	0.197	0.253	0.189	0.552	0.353	0.407	0.313	0.616
RF_ESM	0.269	0.896	0.161	0.579	0.749	0.895	0.645	0.816
GBoost_ESM	0.45	0.849	0.31	0.65	0.761	0.881	0.671	0.827
LogReg_ESM-CDR	0.568	0.525	0.641	0.775	0.708	0.631	0.809	0.862
kSVC_ESM-CDR	0.288	0.426	0.253	0.595	0.48	0.558	0.422	0.681
RF_ESM-CDR	0.372	0.969	0.235	0.617	0.772	0.902	0.676	0.832
GBoost_ESM-CDR	0.501	0.906	0.352	0.673	0.78	0.891	0.695	0.84
LogReg_Antiberty	0.511	0.447	0.632	0.756	0.658	0.547	0.827	0.852
kSVC_Antiberty	0.302	0.382	0.302	0.611	0.43	0.482	0.391	0.657
RF_Antiberty	0.498	0.918	0.353	0.674	0.779	0.885	0.696	0.84
GBoost_Antiberty	0.564	0.802	0.445	0.714	0.771	0.859	0.7	0.84
LogReg_3-mer	0.532	0.674	0.458	0.713	0.758	0.738	0.78	0.865
kSVC_3-mer	0.434	0.831	0.302	0.646	0.783	0.86	0.72	0.849
RF_3-mer	0.498	0.888	0.357	0.675	**0.795**	0.874	0.729	0.855
GBoost_3-mer	0.466	0.867	0.33	0.662	0.768	0.821	0.724	0.848
RBD_VH_VL								
LogReg_ESM	0.319	0.322	0.344	0.6	0.573	0.49	0.694	0.768
kSVC_ESM	0.162	0.265	0.137	0.517	0.314	0.375	0.272	0.587
RF_ESM	0.024	0.6	0.013	0.506	0.626	0.939	0.47	0.732
GBoost_ESM	0.104	0.608	0.058	0.524	0.652	0.859	0.527	0.754
LogReg_CDRextract	0.326	0.325	0.343	0.601	0.592	0.517	0.695	0.776
LogReg_CDRextract	0.326	0.325	0.343	0.601	0.592	0.517	0.695	0.776
kSVC_CDRextract	0.158	0.239	0.13	0.522	0.356	0.454	0.295	0.609
RF_CDRextract	0.034	0.6	0.018	0.509	0.639	0.935	0.486	0.739
GBoost_CDRextract	0.109	0.507	0.062	0.526	0.648	0.873	0.516	0.75
LogReg_Antiberty	**0.408**	0.377	0.463	0.656	0.581	0.49	0.714	0.776
kSVC_Antiberty	0.257	0.36	0.238	0.56	0.271	0.395	0.206	0.569
RF_Antiberty	0.132	0.915	0.071	0.535	0.651	0.935	0.5	0.746
GBoost_Antiberty	0.209	0.714	0.127	0.556	0.66	0.844	0.544	0.761
LogReg_3-mer	0.311	0.365	0.315	0.591	0.561	0.494	0.648	0.752
_3-mer	0.28	0.748	0.175	0.582	0.648	0.909	0.505	0.747
RF_3-mer	0.185	0.935	0.106	0.552	**0.69**	0.93	0.55	0.77
GBoost_3-mer	0.246	0.565	0.166	0.569	0.677	0.84	0.569	0.772

Table 4: Specificity prediction results based on gene expression - distance threshold for similarity based splits 0.05

Dataset	**Similarity based split**				**random_split**			
OVA	F1	prec	recall	roc auc	F1	prec	recall	roc auc
LogReg_GEX_2000_var	0.762	0.657	0.92	0.931	0.775	0.661	0.938	0.938
kSVC_GEX_2000_var	**0.849**	0.963	0.761	0.879	**0.862**	0.967	0.778	0.887
RF_GEX_2000_var	0.267	1	0.157	0.579	0.25	1	0.143	0.572
GBoost_GEX_2000_var	0.815	0.982	0.696	0.847	0.808	0.97	0.695	0.846
LogReg_3-mer	0.519	0.428	0.681	0.785	0.661	0.58	0.778	0.852
kSVC_3-mer	0.574	0.846	0.453	0.722	0.733	0.862	0.638	0.812
RF_3-mer	0.54	0.914	0.421	0.708	0.794	0.897	0.714	0.852
GBoost_3-mer	0.504	0.768	0.426	0.704	0.778	0.862	0.708	0.847
LogReg_GEX_3-mer	0.64	0.497	0.913	0.9	0.714	0.577	0.935	0.924
kSVC_GEX_3-mer	0.753	0.936	0.642	0.818	0.835	0.916	0.768	0.879
RF_GEX_3-mer	0.4	0.98	0.279	0.639	0.711	0.962	0.565	0.781
GBoost_GEX_3-mer	0.767	0.988	0.639	0.819	0.854	0.986	0.754	0.876
RBD								
LogReg_GEX_2000_var	0.866	0.786	0.97	0.957	0.88	0.807	0.969	0.955
kSVC_GEX_2000_var	0.934	0.974	0.898	0.947	0.946	0.99	0.906	0.952
RF_GEX_2000_var	0.774	1	0.633	0.816	0.832	1	0.712	0.856
GBoost_GEX_2000_var	**0.939**	0.988	0.895	0.946	**0.958**	1	0.92	0.96
LogReg_3-mer	0.438	0.398	0.498	0.663	0.65	0.592	0.722	0.799
kSVC_3-mer	0.4	0.783	0.275	0.631	0.745	0.927	0.625	0.806
RF_3-mer	0.217	0.772	0.132	0.564	0.767	0.931	0.655	0.821
GBoost_3-mer	0.343	0.743	0.236	0.612	0.751	0.868	0.665	0.82
LogReg_GEX_3-mer	0.781	0.679	0.926	0.914	0.85	0.769	0.951	0.94
kSVC_GEX_3-mer	0.712	0.966	0.578	0.787	0.85	0.983	0.751	0.874
RF_GEX_3-mer	0.601	1	0.433	0.716	0.849	0.992	0.743	0.871
GBoost_GEX_3-mer	0.9	0.96	0.849	0.921	0.941	0.989	0.898	0.948
OVA_RBD integrated								
LogReg_GEX_2000_var	0.255	0.198	0.364	0.554	0.241	0.189	0.333	0.541
kSVC_GEX_2000_var	0.017	0.281	0.009	0.502	0.009	0.114	0.005	0.499
RF_GEX_2000_var	0	0	0	0.5	0	0	0	0.5
GBoost_GEX_2000_var	0.062	0.255	0.036	0.507	0.072	0.279	0.042	0.511
LogReg_3-mer	0.413	0.345	0.515	0.673	0.596	0.506	0.725	0.8
kSVC_3-mer	0.417	0.797	0.288	0.638	0.701	0.882	0.583	0.785
RF_3-mer	0.281	0.876	0.177	0.587	**0.771**	0.926	0.661	0.826
GBoost_3-mer	0.33	0.722	0.228	0.607	0.747	0.861	0.66	0.82
LogReg_GEX_3-mer	**0.466**	0.382	0.599	0.716	0.599	0.494	0.759	0.811
kSVC_GEX_3-mer	0.378	0.783	0.254	0.621	0.691	0.874	0.573	0.779
RF_GEX_3-mer	0.185	0.867	0.111	0.555	0.659	0.937	0.508	0.751
GBoost_GEX_3-mer	0.318	0.639	0.221	0.601	0.691	0.844	0.585	0.783

A.10 Statistical testing for model comparison

In order to assess the statistical significance of differences in performance between the ML models, we conducted pairwise comparisons using Welch's t-test. The F1 scores were averaged across different train-test splits, features, and model types, with both the mean values and standard deviations presented in Figure 6 and Figure 7 for the model results of antibody sequence features and GEX features, respectively. To correct for multiple comparisons, we applied the Bonferroni adjustment method to the p-values obtained from these tests. In Figure 6 and 7, significance levels are indicated by brackets. ”*”: p <0.05, ”**”: p <0.01, ”***”: p <0.001; the absence indicates non-significant difference.

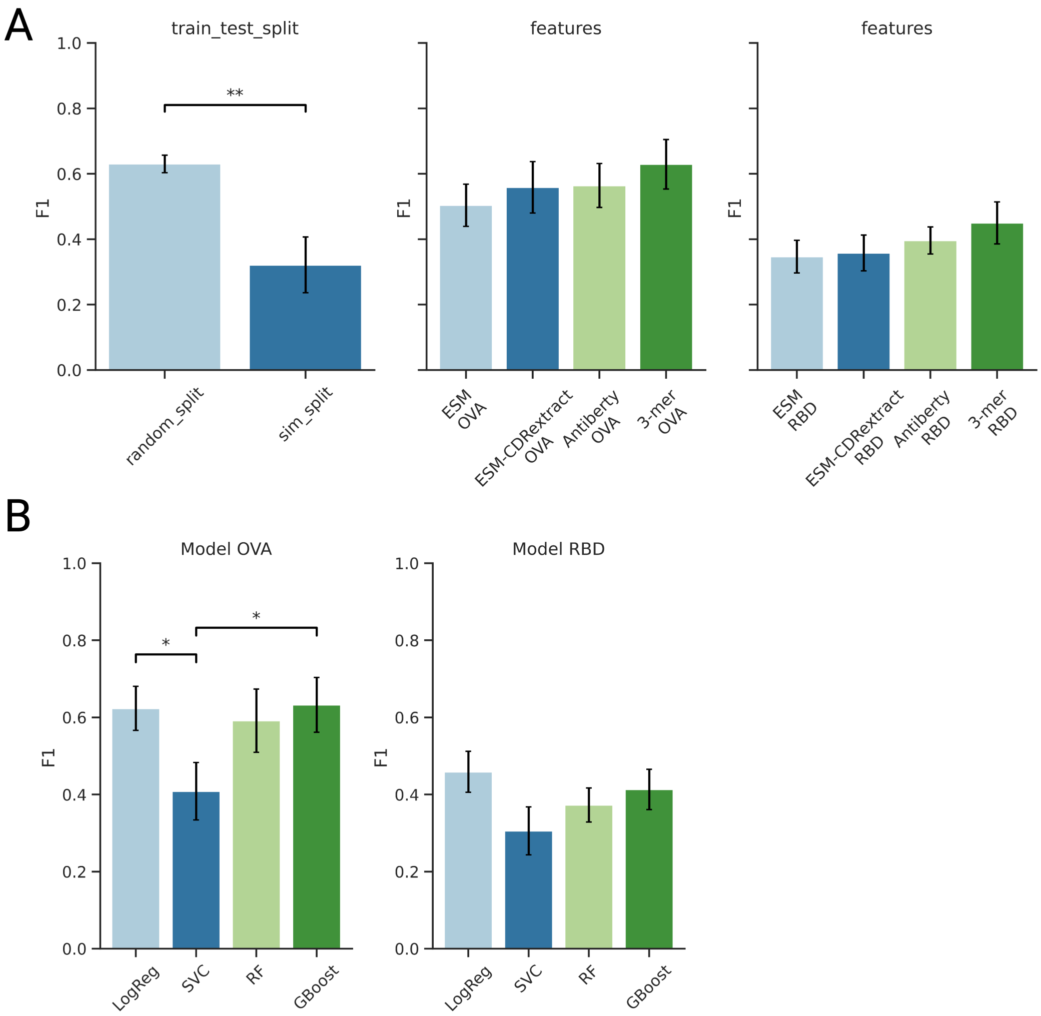

Figure 6: Barplots of averaged antibody sequence model performance for data splitting strategies, features and model types. Significance of pairwise comparisons are indicated with brackets and significance levels are depicted as asterisk ”*”: p <0.05, ”**”: p <0.01, ”***”: p <0.001; If no bracket is drawn, comparisons are not significant.

A.11 Specificity predictions - Evaluation of distance threshold in similarity based train test splitting

To evaluate the impact of the distance threshold utilized to control similarity based train test splits, models were also evaluated using a threshold of 0.1, increasing stringency of sequence similarity splits. Results for predictions based on sequence feature are summarized in Table A.11. See Table 6 for specificity predictions of VH features and refer to Table 8 for GEX feature specificity predictions with train test splits based on distance threshold 0.1.

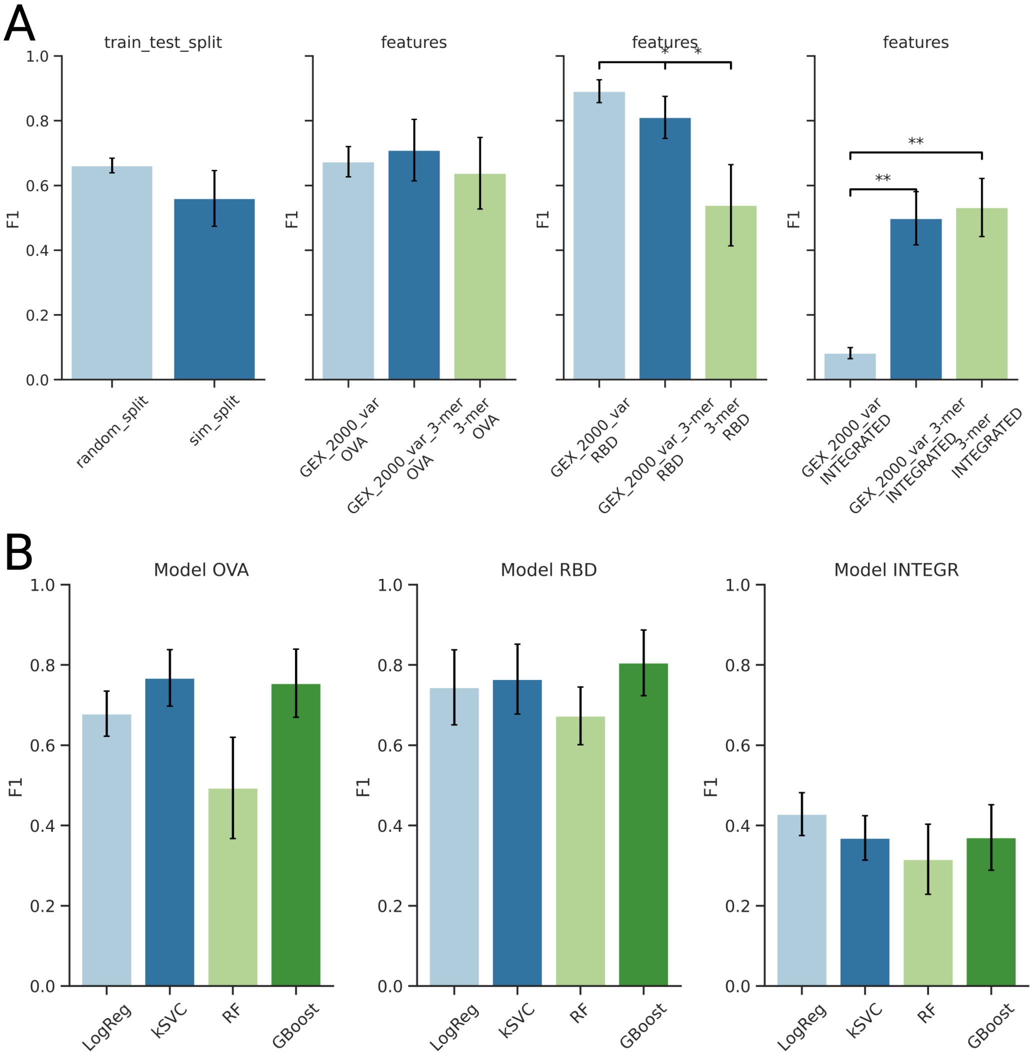

Figure 7: Barplots of aggregated GEX model performance for data splitting strategies, features and model types.Significance of pairwise comparisons are indicated with brackets and significance levels are depicted as asterisk ”*”: p <0.05, ”**”: p <0.01, ”***”: p <0.001; If no bracket is drawn, comparisons are not significant.

A.12 Analysis of coefficients of LogReg GEX models

To further investigate the observed divergence in GEX ML model performance between the separate OVA and RBD datasets versus the integrated dataset, we analyzed the coefficients values learned by the LogReg models. Specifically, we compared the coefficients for genes when the LogReg models were trained on the OVA and RBD datasets individually. We focused on the top 50 genes that exhibited the highest and lowest coefficients in either of the datasets and illustrated the overlap in Figure 8. Notably, there was an overlap of only 3 and 6 genes between those with the smallest and largest coefficients for the OVA and RBD datasets, suggesting divergent patterns learned by the LogReg models across these datasets. Additionally, when comparing the overlap of the integrated dataset with the separate datasets, again only 3-5 genes were common among those with the top and lowest coefficients. Furthermore, the correlation between the coefficients of the genes from the OVA and RBD datasets was minor, with a pearson correlation coefficient of 0.277 (Figure 9). While these findings support our hypothesis regarding the presence of batch effects or distinct antigen-

Table 5: Specificity prediction results based on antibody sequence features - VH_VL - distance split threshold: 0.1

Dataset	**Similarity based split**				**random_split**			
OVA_VH_VL	F1	prec	rec	roc auc	F1	prec	rec	ROC AUC
LogReg_ESM	0.346	0.356	0.39	0.635	0.682	0.596	0.8	0.851
kSVC_ESM	0.198	0.282	0.191	0.552	0.353	0.407	0.313	0.616
RF_ESM	0	0	0	0.5	0.749	0.895	0.645	0.816
GBoost_ESM	0.052	0.45	0.029	0.513	0.761	0.881	0.671	0.827
LogReg_ESM-CDRextract	0.362	0.35	0.397	0.637	0.708	0.631	0.809	0.862
kSVC_ESM-CDRextract	0.182	0.229	0.185	0.558	0.48	0.558	0.422	0.681
RF_ESM-CDRextract	0.014	0.2	0.007	0.504	0.772	0.902	0.676	0.832
GBoost_ESM-CDRextract	0.085	0.32	0.049	0.522	0.78	0.891	0.695	0.84
LogReg_Antiberty	**0.422**	0.366	0.555	0.697	0.658	0.547	0.827	0.852
kSVC_Antiberty	0.225	0.256	0.254	0.574	0.43	0.482	0.391	0.657
RF_Antiberty	0.102	0.334	0.06	0.529	0.779	0.885	0.696	0.84
GBoost_Antiberty	0.135	0.449	0.099	0.546	0.771	0.859	0.7	0.84
LogReg_3-mer	0.218	0.34	0.18	0.552	0.758	0.738	0.78	0.865
kSVC_3-mer	0.106	0.477	0.072	0.532	0.783	0.86	0.72	0.849
RF_3-mer	0.014	0.2	0.007	0.504	**0.795**	0.874	0.729	0.855
GBoost_3-mer	0.082	0.271	0.05	0.514	0.768	0.821	0.724	0.848
RBD_VH_VL	F1	precision	recall	ROC AUC	F1	precision	recall	ROC AUC
LogReg_ESM	0.228	0.227	0.236	0.538	0.573	0.49	0.694	0.768
kSVC_ESM	0.067	0.077	0.08	0.472	0.314	0.375	0.272	0.587
RF_ESM	0	0	0	0.5	0.626	0.939	0.47	0.732
GBoost_ESM	0.026	0.24	0.014	0.504	0.652	0.859	0.527	0.754
LogReg_ESM-CDRextract	0.263	0.289	0.26	0.562	0.592	0.517	0.695	0.776
kSVC_ESM-CDRextract	0.091	0.143	0.082	0.504	0.356	0.454	0.295	0.609
RF_ESM-CDRextract	0	0	0	0.5	0.639	0.935	0.486	0.739
GBoost_ESM-CDRextract	0.018	0.31	0.009	0.502	0.648	0.873	0.516	0.75
LogReg_Antiberty	0.317	0.274	0.383	0.596	0.581	0.49	0.714	0.776
kSVC_Antiberty	0.237	0.279	0.263	0.56	0.271	0.395	0.206	0.569
RF_Antiberty	0.031	0.6	0.016	0.508	0.651	0.935	0.5	0.746
GBoost_Antiberty	0.147	0.56	0.097	0.525	0.66	0.844	0.544	0.761
LogReg_3-mer	0.321	0.298	0.353	0.594	0.561	0.494	0.648	0.752
kSVC_3-mer	0.13	0.533	0.077	0.533	0.648	0.909	0.505	0.747
RF_3-mer	0.008	0.229	0.004	0.501	**0.69**	0.93	0.55	0.77
GBoost_3-mer	0.146	0.375	0.092	0.534	0.677	0.84	0.569	0.772

specific gene signatures Nevertheless, further detailed evaluations are necessary to fully validate these observations.

A.13 Classification based on differentially expressed genes

To ascertain the utility and necessity of ML methods in predicting antigen specificity from GEX data, we designed a control experiment utilizing the top 10 DE genes, both upregulated and downregulated. First, the integrated OVA_RBD dataset was divided into a 70% training and 30% test split. DESeq2 was utilized on pseudo-bulk gene expression data (Love et al., 2014; Squair et al., 2021) to identify DE genes in the 70% training dataset. Subsequently, gene module scores were computed for each cell, defined as the mean expression level of the identified genes, adjusted by the average expression of randomly selected control gene sets. We calculated the mean of the module scores of the antigen-specific cell fraction in the training dataset, which was used to classify cells in the test dataset. Cells were deemed specific, if the gene module score for the upregulated genes was at or above and the score for the downregulated genes was at or below the respective mean scores determined from the training set. The outcomes of this experiment are detailed in Table 9. Summarized results reveal that this classification approach outperformed models using the top 2000

Table 6: Specificity prediction evaluation results based on antibody sequence features from VH only - distance split threshold: 0.05

Dataset	**Similarity based split**				**random_split**			
OVA_VH	F1	prec	recall	roc auc	F1	prec	recall	roc auc
LogReg_ESM	0.561	0.498	0.701	0.779	0.667	0.56	0.824	0.855
kSVC_ESM	0.333	0.529	0.255	0.608	0.481	0.664	0.377	0.672
RF_ESM	0.275	0.852	0.175	0.584	0.69	0.884	0.569	0.778
GBoost_ESM	0.419	0.714	0.307	0.644	0.732	0.866	0.637	0.809
LogReg_CDRextract	0.54	0.497	0.646	0.76	0.686	0.583	0.835	0.864
kSVC_CDRextract	0.38	0.565	0.307	0.631	0.565	0.684	0.482	0.721
RF_CDRextract	0.344	0.862	0.238	0.615	0.734	0.914	0.616	0.803
GBoost_CDRextract	0.454	0.762	0.351	0.665	0.757	0.87	0.673	0.828
LogReg_Antiberty	0.472	0.373	0.681	0.739	0.616	0.491	0.83	0.838
kSVC_Antiberty	0.377	0.499	0.319	0.628	0.498	0.649	0.405	0.683
RF_Antiberty	0.434	0.745	0.332	0.656	0.709	0.834	0.619	0.799
GBoost_Antiberty	0.46	0.688	0.365	0.668	0.713	0.811	0.637	0.805
LogReg_3-mer	**0.619**	0.651	0.642	0.787	0.731	0.679	0.793	0.863
kSVC_3-mer	0.493	0.783	0.375	0.678	0.757	0.844	0.69	0.834
RF_3-mer	0.487	0.775	0.38	0.68	**0.781**	0.844	0.73	0.853
GBoost_3-mer	0.538	0.79	0.423	0.701	0.769	0.836	0.715	0.845
RBD_VH								
LogReg_ESM	**0.388**	0.372	0.432	0.637	0.572	0.495	0.678	0.764
kSVC_ESM	0.162	0.247	0.136	0.53	0.341	0.443	0.278	0.601
RF_ESM	0.071	0.782	0.038	0.516	0.611	0.922	0.458	0.725
GBoost_ESM	0.158	0.634	0.1	0.541	0.625	0.836	0.5	0.739
LogReg_CDRextract	**0.388**	0.371	0.431	0.636	0.575	0.491	0.694	0.769
kSVC_CDRextract	0.181	0.338	0.168	0.534	0.394	0.541	0.309	0.626
RF_CDRextract	0.07	0.967	0.037	0.518	0.633	0.941	0.478	0.736
GBoost_CDRextract	0.198	0.719	0.121	0.557	0.654	0.876	0.523	0.753
LogReg_Antiberty	0.385	0.361	0.435	0.64	0.574	0.484	0.708	0.772
kSVC_Antiberty	0.224	0.437	0.22	0.569	0.428	0.615	0.33	0.642
RF_Antiberty	0.272	0.73	0.198	0.586	0.64	0.917	0.492	0.741
GBoost_Antiberty	0.26	0.379	0.216	0.577	0.625	0.815	0.508	0.741
LogReg_3-mer	0.302	0.265	0.395	0.583	0.532	0.459	0.631	0.735
kSVC_3-mer	0.243	0.636	0.192	0.578	0.633	0.903	0.488	0.738
RF_3-mer	0.215	0.788	0.169	0.572	0.673	0.897	0.539	0.763
GBoost_3-mer	0.218	0.414	0.182	0.56	0.671	0.831	0.562	0.769

variable genes via LogReg or kSVC with F1 score of 0.3968. However, this method still fell short of the performance achieved by LogReg models integrating GEX and antibody sequence information (3-mer sequences).

Table 7: Specificity prediction evaluation results based on antibody sequence features from VH only - distance split threshold: 0.1

Dataset	**Similarity based split**				**random_split**			
OVA_VH	F1	prec	recall	roc auc	F1	prec	recall	roc auc
LogReg_ESM	0.316	0.335	0.382	0.619	0.667	0.56	0.824	0.855
kSVC_ESM	0.234	0.618	0.174	0.579	0.481	0.664	0.377	0.672
RF_ESM	0.133	0.2	0.1	0.55	0.69	0.884	0.569	0.778
GBoost_ESM	0.16	0.425	0.102	0.547	0.732	0.866	0.637	0.809
LogReg_ESM-CDRextract	0.31	0.394	0.342	0.621	0.686	0.583	0.835	0.864
kSVC_ESM-CDRextract	0.316	0.468	0.3	0.63	0.565	0.684	0.482	0.721
RF_ESM-CDRextract	0.005	0.2	0.002	0.5	0.734	0.914	0.616	0.803
GBoost_ESM-CDRextract	0.137	0.414	0.085	0.538	0.757	0.87	0.673	0.828
LogReg_Antiberty	**0.37**	0.378	0.38	0.633	0.616	0.491	0.83	0.838
kSVC_Antiberty	0.251	0.393	0.197	0.576	0.498	0.649	0.405	0.683
RF_Antiberty	0.157	0.533	0.112	0.556	0.709	0.834	0.619	0.799
GBoost_Antiberty	0.195	0.327	0.153	0.557	0.713	0.811	0.637	0.805
LogReg_3-mer	0.357	0.49	0.358	0.646	0.731	0.679	0.793	0.863
kSVC_3-mer	0.291	0.524	0.268	0.616	0.757	0.844	0.69	0.834
RF_3-mer	0.08	0.2	0.05	0.525	**0.781**	0.844	0.73	0.853
GBoost_3-mer	0.153	0.387	0.096	0.539	0.769	0.836	0.715	0.845
RBD_VH								
LogReg_ESM	0.176	0.181	0.217	0.533	0.572	0.495	0.678	0.764
kSVC_ESM	0.177	0.229	0.159	0.547	0.341	0.443	0.278	0.601
RF_ESM	0	0	0	0.496	0.611	0.922	0.458	0.725
GBoost_ESM	0.055	0.097	0.04	0.488	0.625	0.836	0.5	0.739
LogReg_ESM-CDRextract	0.192	0.194	0.219	0.519	0.575	0.491	0.694	0.769
kSVC_ESM-CDRextract	0.166	0.226	0.242	0.562	0.394	0.541	0.309	0.626
RF_ESM-CDRextract	0	0	0	0.497	0.633	0.941	0.478	0.736
GBoost_ESM-CDRextract	0.022	0.083	0.013	0.498	0.654	0.876	0.523	0.753
LogReg_Antiberty	0.177	0.277	0.187	0.551	0.574	0.484	0.708	0.772
kSVC_Antiberty	0.136	0.138	0.188	0.532	0.428	0.615	0.33	0.642
RF_Antiberty	0	0	0	0.498	0.64	0.917	0.492	0.741
GBoost_Antiberty	0.087	0.126	0.069	0.51	0.625	0.815	0.508	0.741
LogReg_3-mer	0.198	0.165	0.272	0.475	0.532	0.459	0.631	0.735
kSVC_3-mer	0.024	0.308	0.012	0.5	0.633	0.903	0.488	0.738
RF_3-mer	0	0	0	0.499	**0.673**	0.897	0.539	0.763
GBoost_3-mer	0.038	0.13	0.026	0.482	0.671	0.831	0.562	0.769

Table 8: Specificity prediction results based on gene expression - sequence distance split 0.1 based on antibody sequences

Dataset	**Similarity based split**				**random_split**			
OVA	F1	prec	recall	roc auc	F1	prec	recall	roc auc
LogReg_GEX_2000_var	0.707	0.595	0.902	0.918	0.775	0.661	0.938	0.938
kSVC_GEX_2000_var	**0.815**	0.951	0.717	0.857	**0.862**	0.967	0.778	0.887
RF_GEX_2000_var	0.273	1	0.165	0.582	0.25	1	0.143	0.572
GBoost_GEX_2000_var	0.795	0.966	0.678	0.838	0.808	0.97	0.695	0.846
LogReg_3-mer	0.323	0.299	0.372	0.635	0.661	0.58	0.778	0.852
kSVC_3-mer	0.21	0.559	0.132	0.561	0.733	0.862	0.638	0.812
RF_3-mer	0.007	0.2	0.003	0.502	0.794	0.897	0.714	0.852
GBoost_3-mer	0.084	0.771	0.047	0.523	0.778	0.862	0.708	0.847
LogReg_GEX_2000_var_3-mer	0.577	0.439	0.868	0.871	0.714	0.577	0.935	0.924
kSVC_GEX_2000_var_3-mer	0.441	0.853	0.313	0.654	0.835	0.916	0.768	0.879
RF_GEX_2000_var_3-mer	0.046	0.6	0.024	0.512	0.711	0.962	0.565	0.781
GBoost_GEX_2000_var_3-mer	0.512	0.986	0.348	0.674	0.854	0.986	0.754	0.876
RBD								
LogReg_GEX_2000_var	0.85	0.764	0.966	0.952	0.88	0.807	0.969	0.955
kSVC_GEX_2000_var	**0.935**	0.978	0.897	0.947	0.946	0.99	0.906	0.952
RF_GEX_2000_var	0.807	1	0.68	0.84	0.832	1	0.712	0.856
GBoost_GEX_2000_var	0.947	0.992	0.907	0.953	**0.958**	1	0.92	0.96
LogReg_3-mer	0.389	0.36	0.451	0.631	0.65	0.592	0.722	0.799
kSVC_3-mer	0.426	0.831	0.297	0.642	0.745	0.927	0.625	0.806
RF_3-mer	0.099	0.383	0.059	0.527	0.767	0.931	0.655	0.821
GBoost_3-mer	0.211	0.37	0.151	0.555	0.751	0.868	0.665	0.82
LogReg_GEX_2000_var_3-mer	0.723	0.611	0.92	0.89	0.85	0.769	0.951	0.94
kSVC_GEX_2000_var_3-mer	0.707	0.973	0.566	0.781	0.85	0.983	0.751	0.874
RF_GEX_2000_var_3-mer	0.612	1	0.447	0.724	0.849	0.992	0.743	0.871
GBoost_GEX_2000_var_3-mer	0.917	0.98	0.862	0.929	0.941	0.989	0.898	0.948
OVA_RBD integrated								
LogReg_GEX_2000_var	0.218	0.18	0.302	0.527	0.241	0.189	0.333	0.541
kSVC_GEX_2000_var	0.009	0.213	0.005	0.5	0.009	0.114	0.005	0.499
RF_GEX_2000_var	0	0	0	0.5	0	0	0	0.5
GBoost_GEX_2000_var	0.037	0.203	0.021	0.501	0.072	0.279	0.042	0.511
LogReg_3-mer	0.378	0.335	0.449	0.641	0.596	0.506	0.725	0.8
kSVC_3-mer	0.324	0.704	0.216	0.602	0.701	0.882	0.583	0.785
RF_3-mer	0.02	0.56	0.01	0.505	**0.771**	0.926	0.661	0.826
GBoost_3-mer	0.151	0.566	0.091	0.537	0.747	0.861	0.66	0.82
LogReg_GEX_2000_var_3-mer	**0.437**	0.367	0.55	0.689	0.599	0.494	0.759	0.811
kSVC_GEX_2000_var_3-mer	0.241	0.698	0.148	0.569	0.691	0.874	0.573	0.779
RF_GEX_2000_var_3-mer	0.018	0.467	0.009	0.504	0.659	0.937	0.508	0.751
GBoost_GEX_2000_var_3-mer	0.155	0.513	0.093	0.538	0.691	0.844	0.585	0.783

Table 9: Specificity prediction results based on gene expression - sequence distance split 0.05 based on antibody sequence sequences

DE genes	**F1**	**prec**	**recall**
Top10_down	0.3084	0.2049	0.6229
Top10_up	0.3825	0.3227	0.4695
Top10_down & Top10_up	0.3968	0.3805	0.4144

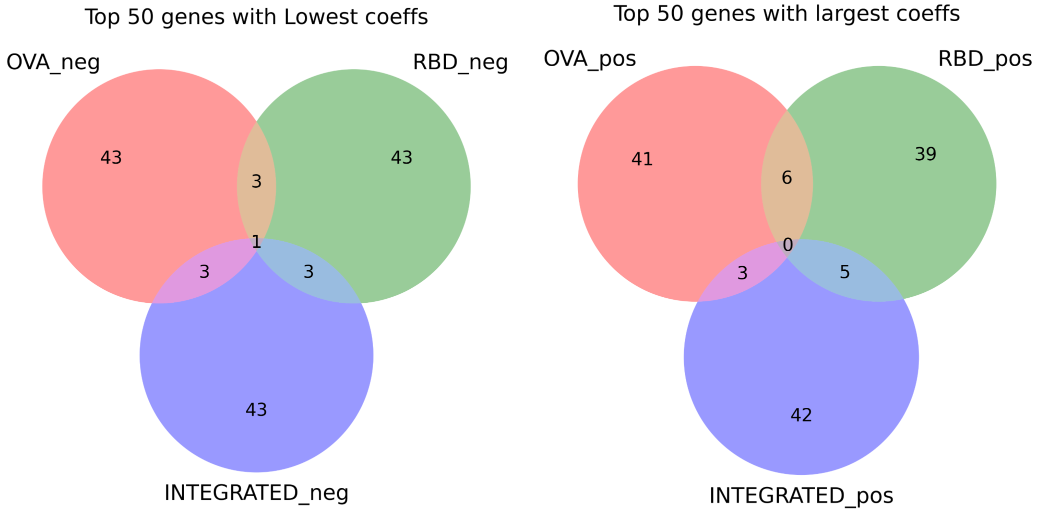

Figure 8: Overlap of genes within the top 50 highest and lowest coefficients from LogReg models trained on the OVA and RBD dataset.

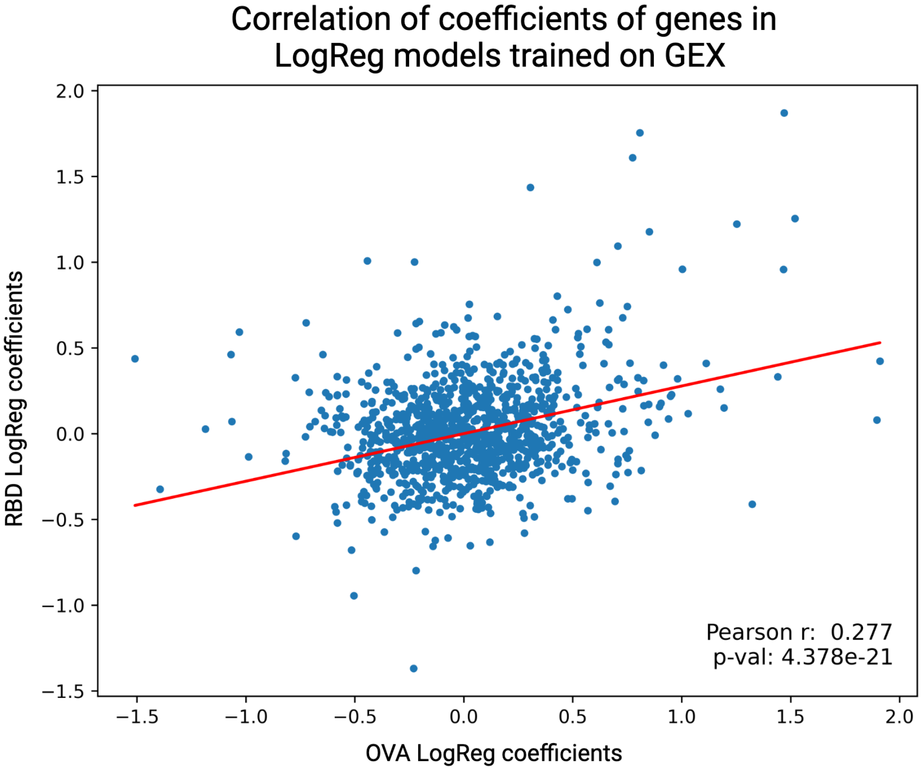

Figure 9: Correlation plot of the values of the coefficients from LogReg models trained on OVA and the RBD GEX features.