Using causal modeling to analyze generalization of biomarkers in high-dimensional domains: a case study of adaptive immune repertoires

Milena Pavlović [1] **Ghadi S. Al Hajj** [1] **Victor Greiff** [2] **Johan Pensar** [3] **Geir Kjetil Sandve** [1]

Abstract

Machine learning is increasingly used to discover diagnostic and prognostic biomarkers from high-dimensional molecular data. However, a variety of factors related to experimental design may affect the ability to learn generalizable and clinically applicable diagnostics. Here, we discuss building a diagnostic based on a specific, recently established high-dimensional biomarker – adaptive immune receptor repertoires (AIRRs), and investigate how causal modeling may improve the robustness and generalization of developed diagnostics. We examine how the main biological and experimental factors of the AIRR domain may influence the learned biomarkers, especially in the presence of dataset shifts, and provide simulations of such effects. We conclude that causal modeling could improve AIRR-based diagnostics, but also that causal modeling itself might find a powerful testbed with complex, high-dimensional variables in the AIRR field.

1. Introduction

High-throughput sequencing technologies now allow for the examination of a variety of patient characteristics (Frazer et al., 2009; Locke et al., 2019; Byron et al., 2016; Huang et al., 2021; Arnaout et al., 2021; Greiff et al., 2020). Proof-of-concept studies showed that such molecular markers hold great promise for disease diagnostics, especially in combination with machine learning (ML) (Locke et al., 2019; Byron et al., 2016; Arnaout et al., 2021; Maros et al., 2020). However, there exist several challenges to using ML for diagnostics. First, the data used in diagnostic studies may be selected based on availability, e.g., collected from patients visiting the clinic or having a similar genetic background (sometimes referred to as ”convenience sampling”). Furthermore, the data might be collected at multiple locations or at distinct time points. These factors may introduce systematic differences between datasets, such as batch effects, which need to be taken into account when designing a new study, or adjusted for when the data are already collected. A failure to do so can lead to models failing in real-world application despite showing promising performance during diagnostic development (Subbaswamy & Saria, 2020; Castro et al., 2020; Whalen et al., 2021; Dockès et al., 2021). Finally, biomarker data are typically high-dimensional, which makes it more challenging to disentangle noise, biases, and other conditions from the true markers associated with the disease (Teschendorff, 2019; Weber et al., 2022).

Addressing the biases that emerge due to environment change, confounding, and sample selection are main challenges in developing ML-based diagnostics. The causal inference framework described by Pearl (2009) can be used to estimate causal effects from non-experimental data whenever the effect is identifiable under a given causal structure, and could help in resolving these biases. More specifically, a causal model of the biological process may determine which mechanisms (conditional distributions of variables of interest given their direct causes) are likely to remain stable across populations. The ML model may then be based on these invariant mechanisms (Schölkopf et al., 2021) ensuring the robustness of the ML model in the presence of different dataset shifts (Subbaswamy & Saria, 2020; Storkey, 2008; Kouw & Loog, 2018). Additionally, causal models can help to formally or intuitively reason about how diagnostic accuracy may be affected by a variety of differences between application contexts.

In this paper, we focus on one particular use case – that of adaptive immune receptor repertoires (AIRRs). AIRRs are high-dimensional molecular markers reflecting past and present immune responses of a patient and can be efficiently assayed based on targeted high-throughput sequencing from a standard blood sample. These approaches may enable earlier diagnosis, complement existing diagnostic tests, and have in principle the capacity to diagnose a broad range of diseases by a single test (Arnaout et al., 2021). For this rea-

[1] Department of Informatics, University of Oslo, Oslo, Norway [2] Department of Immunology, University of Oslo, Oslo, Norway [3] Department of Mathematics, University of Oslo, Norway. Correspondence to: Geir Kjetil Sandve <geirksa@ifi.uio.no>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

son, AIRRs are highly promising as biomarkers of immune-mediated diseases (Arnaout et al., 2021; Greiff et al., 2020), like cancer (Ostmeyer et al., 2019; Beshnova et al., 2020), celiac disease (Shemesh et al., 2021; Yao et al., 2021), multiple sclerosis (Ostmeyer et al., 2017), rheumatoid arthritis (Liu et al., 2019), systemic lupus erythematosus (Liu et al., 2019), cytomegalovirus (Emerson et al., 2017), and hepatitis C virus (Eliyahu et al., 2018). Here, we discuss why and how accounting for the biological and experimental aspects of the underlying data-generating processes is crucial to avoid capturing spurious correlations and to anticipate changes in diagnostic performance between development and deployment (the clinic).

2. Adaptive immune receptor repertoires capture all immune-related diseases

Adaptive immune receptors (AIR) are proteins used by B and T cells to recognize foreign threats like viruses or cancerous cells and mount an immune response to destroy them. AIRs are generated by an individualized stochastic process (Slabodkin et al., 2021) resulting in an estimated 10^{15} different receptors (Sewell, 2012; Nikolich-Žugich et al., 2004; Zarnitsyna et al., 2013; Murugan et al., 2012), defining an extremely high-dimensional space for biomarker discovery. When examining individual AIRs, one is mostly analyzing an $\sim$15 amino acid long part of the receptor called CDR3.

The set of all AIRs present in an individual is referred to as the adaptive immune receptor repertoire (AIRR). There are an estimated 10^8 unique receptors (Elhanati et al., 2015; Greiff et al., 2015; Elhanati et al., 2018) sampled from the pool of 10^{15} possible receptors, leading to the very low overlap of receptors between AIRRs of different individuals (Elhanati et al., 2018). Additionally, very few receptors in an AIRR are specific to any one disease, and the rules determining disease specificity of receptors are largely unknown due to complex sequence patterns. See the review by Greiff and colleagues (2020) for a general discussion of AIRR ML.

3. Challenges in AIRR diagnostics study design

The typical workflow for building an AIRR-based diagnostic is to select a study cohort of disease affected (cases) and healthy individuals (controls) from an underlying source population, collect blood or tissue samples from each individual, and perform targeted high-throughput DNA sequencing to obtain a set of approximately 10^5–10^6 immune receptors per individual that can be used to learn the patterns indicative of disease (Figure 1). Current diagnostic studies typically recruit in the order of $100 - 1000$ individuals for learning disease-specific AIRR-based biomarkers. The common assumption that training and deployment data come in the form of independent samples from a common underlying distribution (the i.i.d. assumption) rarely holds (Nestor et al., 2018; Ghassemi et al., 2020). Marginal distributions may change due to label shift (e.g., change in disease prevalence) or covariate shift (e.g., change in age distribution). The conditional distribution of variables may change if it describes an anticausal relation (when predicting the cause from the effect, e.g., immune state from AIRR) or due to the occurrence of unstable mechanisms (Subbaswamy & Saria, 2020) (e.g., changing the time of sequencing in the course of the disease might result in estimates that only hold for the study cohort). The biases may arise from different aspects of the data generating process leading to the identification problem as defined in the causal inference (Hernán & Robins, 2020).

To illustrate these biases in the AIRR domain, we introduce an example of building a diagnostic for a viral infection (Figure 2). In this example, the immune state is defined as the presence of the pathogen of interest in an individual in a way that gives rise to changes in AIRR. In addition to the immune state, AIRR is also influenced by prior immune events (e.g., prior infections or vaccinations), age (Britanova et al., 2014), sex (Schneider-Hohendorf et al., 2018), genetics (Slabodkin et al., 2021), and the environment. The human leukocyte antigen (HLA), a genetic component determining the presentation of pathogens to initiate the immune response, influences the AIRR as well either through its type or though level of expression on the cell surface (Dendrou et al., 2018; Ishigaki et al., 2022). Finally, the observed sequencing AIRR data reflect only a limited proportion of a patient's full AIRR, introducing additional sampling variability and sequencing protocol biases (Barennes et al., 2021; Trück et al., 2021). For other types of diseases, such as autoimmunity or cancer, the graphs and connections might differ. We expand on the characteristics and implications of these variables and their relations in the remainder of the paper.

3.1. Representation and dimensionality of AIRR data

Building diagnostics based on AIRRs (and molecular data in general) is made more challenging by the high dimensionality of the data. While we have represented an AIRR by a single node in the causal graph (Figure 2), it represents millions of individual AIRs. How to represent these data for causal or machine learning analysis is not trivial. A typical ML representation would be to consider each possible immune receptor sequence as a feature, annotating each patient with an indicator vector of which sequences are present. As there are more than 10^{15} such receptors, this representation is however not computationally feasible. Alternatively, each AIRR can be represented directly as the set of sequences present in a patient. A challenge with this representation is the lack of a common feature space across patients (ex-

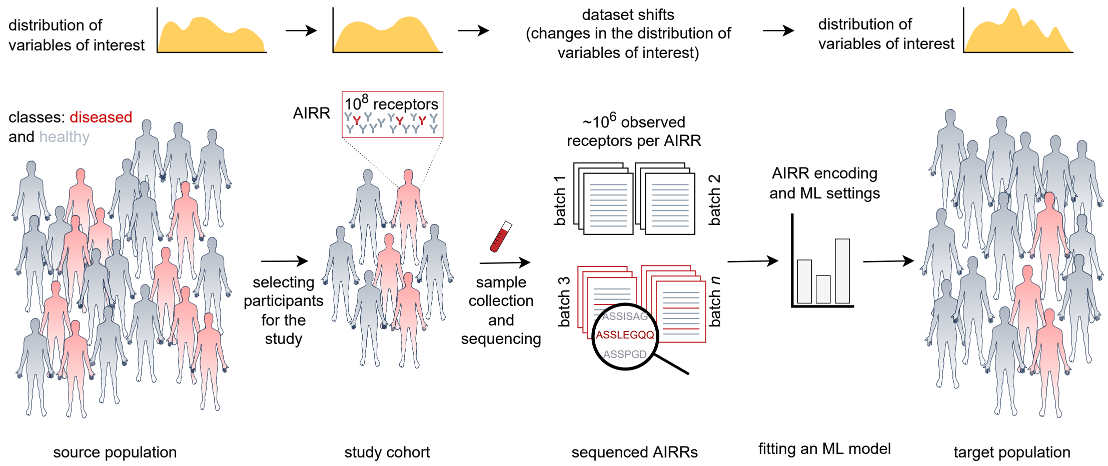

Figure 1. Overview of the AIRR-based diagnostic development pipeline

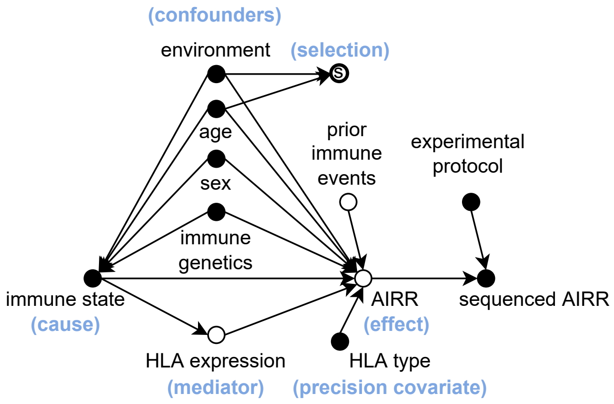

Figure 2. An example of an AIRR causal graph for a viral infection.

amples), which is required by the majority of causal and ML approaches. Some form of aggregation across receptors in a repertoire is thus typically needed. However, even representing a single receptor sequence is not trivial. A basic representation is to one-hot encode each amino acid of the sequence, which leads to a low-dimensional feature space (with 20 different amino acids in $\sim$15 long receptor sequences) with very strong interaction effects between these dimensions since neighboring amino acids together influence the folding of the resulting receptor protein and its binding to the pathogen. To capture at least the low-order interaction effects directly in single dimensions, another common receptor representation is to consider all k-mers for a given k as features, as k-mers are in contact with the pathogen (Akbar et al., 2021; Ostmeyer et al., 2019; Dash et al., 2017), and then represent each receptor as an indicator vector of k-mers being present. Repertoire representations can then also be constructed by aggregating k-mer presence values for all receptors. To avoid manual feature construction, the latent AIRR representation could be learned from the data, e.g., via autoencoders or similar methods coupled with domain-specific restrictions (Widrich et al., 2020; Sidhom et al., 2021; Davidsen et al., 2019), although interpreting the learned representation and associating it with the causal model may be a challenging task.

3.2. Confounder variables

A variety of patient factors, like age (Britanova et al., 2014), sex (Schneider-Hohendorf et al., 2018), and genetic background (Krishna et al., 2020) have been demonstrated to influence the immune repertoire, and are often also associated with disease risk, making them act as confounders in disease diagnostics. While confounding is in general not problematic for predictive purposes, the distribution of these particular factors may be dramatically different between training and deployment settings, e.g., based on convenience of access to patients with particular characteristics during training, as well as various forms of pre-screening before a diagnostic is to be used in a clinical setting. It may thus be important to explicitly consider statistical transportability (Correa & Bareinboim, 2019) of immune-based diagnostics for a particular disease. Additionally, the recovered biomarkers could be reflecting the confounders just as much as the immune state (Whalen et al., 2021) so that if the aim is to obtain biological insight (discover causal effects), confounding should be controlled for. However, due to the complex high-dimensional nature of AIRRs, it is very hard to learn the relations between AIRRs and patient factors, restricting the ability to perform reasonably precise confounder correction or analysis of sensitivity to distributional shifts of patient characteristics.

3.3. Measurement timing and batch effects

Batch effects are systematic biases in the statistical sense, connected to experimental protocols exhibiting different behavior across conditions (Leek et al., 2010). Although always present, the batch effects are more problematic when correlated with the label (e.g., immune state), for which a predictive model may achieve good performance in a study cohort by learning batch effect associations but fail to generalize to deployment settings. Additionally, sequencing errors in the AIRR domain (Barennes et al., 2021) are challenging to correct since the ground truth (e.g., exact receptor sequences) is not known.

The timing of sequencing in the course of the disease is also very important for diagnostic development. For example, the positive examples (diseased individuals) in the dataset used for training might be collected retrospectively, after individuals exhibited symptoms of the disease and were diagnosed and possibly already treated by medical professionals. In this case, the collected AIRRs will not be representative of the AIRRs of individuals who would get tested to establish the diagnosis. To mitigate this issue, the study cohort should be representative of the target population in terms of the timing of measurement or at least include individuals sequenced across the disease progression spectrum.

3.4. HLA – one variable, many roles

Some biological variables can have different roles in the causal graph depending on the disease or immune event, with important implications for the analysis. One such example is the genetic variants that a particular individual has for a molecule known as human leukocyte antigen (HLA). For adaptive immune cells to recognize a pathogen, fragments derived from the pathogen must be presented on the surface of a cell, bound to an HLA molecule. Genetic variation of the HLA molecule influences how a given peptide is presented, which again influences which immune receptors are recognizing the pathogen. Due to the way newly created immune cells are filtered during their development (known as thymic selection), the HLA type will have some influence on the overall distribution of immune receptor sequences. For a viral infection, HLA type is not assumed to influence disease risk and thus has the role of a precision covariate. However, for many autoimmune diseases HLA is known to also influence disease risk, and thus has the role of a confounder. Additionally, HLA can be a moderator of how given immune receptors are contributing to disease risk. For example, the presence of particular HLA types is a necessary condition for potentially harmful immune cells to be causing celiac disease. Finally, in some cancers, tumor cells have mutations that affect the functioning of HLA and help tumor cells evade immune recognition (Schaafsma et al., 2021). In this case, HLA acts as a mediator between disease and AIRR.

The different roles of HLA also point to different strategies potentially being useful. In diseases where HLA type acts as a confounder, a diagnostic model might exploit the backdoor path from repertoire via HLA to disease risk. For predictive purposes, this would primarily contribute to improved diagnostic accuracy, although the accuracy observed during training may then not be representative of deployment settings where the HLA distribution is very different. This distributional difference is common, as patients to be diagnosed may be pre-selected by a separate test of HLA type. In diseases where HLA type acts as a moderator, it may be useful to at least analyze predictive accuracy stratified by HLA type, and potentially also to enrich for patients with challenging HLA types in study recruitment, so as to have more available data for learning more subtle predictive patterns for this stratum. As also discussed in the section on confounders, even though the existence of a dependency between AIRR and HLA is biologically highly plausible, the complex nature of AIRRs makes it very hard to estimate the precise form of the relation from available patient data.

4. Experiments show the influence of the underlying causality on predictive models

To illustrate the influence of different variables in the causal model on the performance of ML algorithms, we perform two experiments where we train the algorithm to predict the immune state from the AIRR data without considering potential biases ((Figure 3)). The first experiment illustrates how the confounding influences the prediction of the immune state (Figure 3a-c). The confounder and immune state are binary variables: the confounder has values C1 or C2 and the immune state can be diseased (positive) or healthy (negative). The parameters of the probability distribution of the immune state depend on the value of the confounder, making examples with confounder C1 much more likely to be diseased. We constrain the resulting distribution of the immune state variable to have balanced classes (approximately the same number of diseased and healthy examples), following a typical setting in ML, and then vary the confounder distribution while keeping the classes balanced to examine the effect of the confounder on the prediction performance. As discussed previously, the presence of a confounder is not always an issue for the prediction task: if the confounder distribution does not change from source to target population, similar performance can be expected on both datasets (Figure 3b). However, if the confounder distribution changes, the performance may drop (Figure 3c).

In the second experiment (Figure 3d), we examine how selection bias and batch effects influence immune state prediction. The causal graph includes the immune state which causes changes in AIRRs, the hospital the patients came

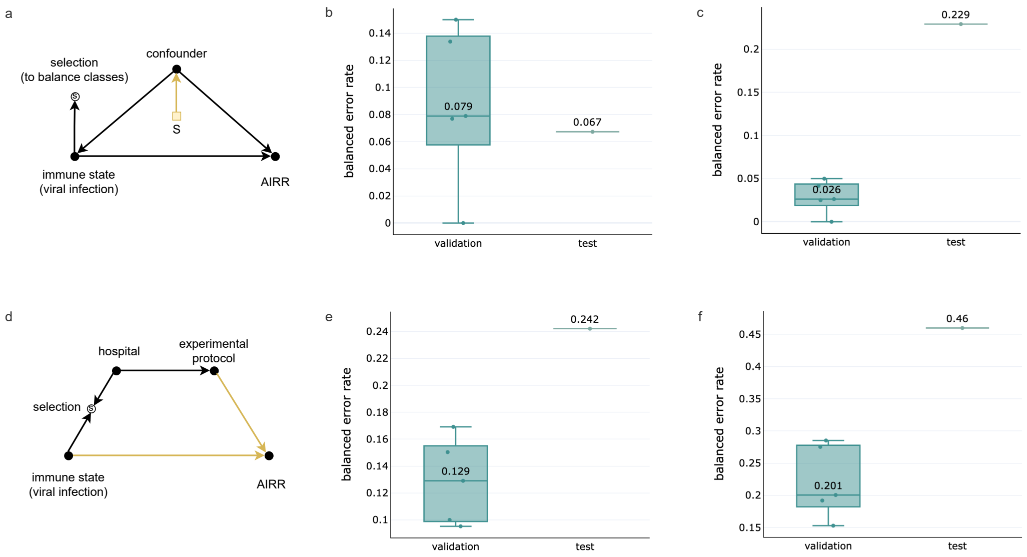

Figure 3. Experiments showing the influence of different variables from the AIRR causal model on immune state prediction task.

from, the experimental protocol for sequencing AIRRs set by the hospital, and the sequenced AIRRs themselves. We simulate the training data in the presence of selection bias, which introduces a correlation between hospital and immune state, even though the hospital does not have any influence on it. When an ML approach is trained on data that are biased in this way, it also learns the signal of the sequencing protocol since it is predictive of the immune state through the spurious correlation between the immune state and the hospital. When such a model is then applied to new data that are not biased, its performance decreases (Figure 3e).

We also show that even in the absence of any relation between the immune state and AIRR, it is possible for an ML model to learn a spurious correlation as a consequence of the selection bias and achieve deceivingly good performance, depending on how strong the correlation is between the immune state and protocol (Figure 3f).

For experiments, we simulated 200 AIRRs for training and 100 for testing using OLGA (Sethna et al., 2019) for AIRR simulations with 500 TCR sequences per AIRR from the default human OLGA model. We used DagSim (Al Hajj et al., 2022) to simulate the data from the causal graph. As the next step, we implanted 3-mers into some of the AIRs of individual repertoires with the appropriate immune state and confounder values. To assess ML performance, we encoded the AIRRs via 3-mer frequencies and fitted logistic regression with the L1 penalty to perform the prediction, as previously described by (Kanduri et al., 2022). The model performance was measured via balanced error rate. For AIRR-ML analyses, we used immuneML (Pavlović et al., 2021). All analyses are available at https://github.com/uio-bmi/CausalAIRR.

5. Conclusion

There has recently been substantial theoretical development on how underlying causality influences the stability of predictive ML models. An interesting application area for this theory is to help anticipate the behavior and performance of diagnostics when translated from a study setting to clinical application. To materialize this potential for practical application, we believe that the field should work to increasingly consider the broad variety of subtleties and complexities inherent to real scenarios. We here proposed AIRR to be a particularly interesting case. In our understanding, the methodology for e.g., analyzing path contributions and sensitivity to distributional shifts is still in its infancy for settings involving complex and high-dimensional variables like AIRRs. Since the AIRR setting offers a clear mechanistic understanding of several aspects of the underlying causality having a basis in immune cells with discrete disease-specific receptor sequences, we believe it can provide a powerful testbed for the development of new methodology for estimating causal relations that involve complex variables.

An additional observation from our use case is that a causal analysis of immune-based diagnostics might need to consider subtly different causal models for each disease in question. Even the causal direction between AIRRs and immune state may need to be modeled differently across diseases, where immune cells can be seen as causing autoimmune disease, while a viral infection is more naturally seen as causing an immune response. Also, the same variable can play different roles in different disease settings, where e.g., HLA can be seen as either a precision covariate, a confounder, and/or a moderator depending on the disease.

References

Akbar, R., Robert, P. A., Pavlović, M., Jeliazkov, J. R., Snapkov, I., Slabodkin, A., Weber, C. R., Scheffer, L., Miho, E., Haff, I. H., Haug, D. T. T., Lund-Johansen, F., Safonova, Y., Sandve, G. K., and Greiff, V. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, 34(11):108856, March 2021. ISSN 2211-1247. doi: 10.1016/j.celrep.2021.108856.

Al Hajj, G. S., Pensar, J., and Sandve, G. K. DagSim: Combining DAG-based model structure with unconstrained data types and relations for flexible, transparent, and modularized data simulation, May 2022. URL http://arxiv.org/abs/2205.11234. Number: arXiv:2205.11234 arXiv:2205.11234 [cs].

Arnaout, R. A., Prak, E. T. L., Schwab, N., Rubelt, F., Community, t. A. I. R. R., Arnaout, R. A., Arora, R., Bashford-Rogers, R., Breden, F., Bukhari, S. A. C., Corrie, B., Cowell, L. G., Efroni, S., Gooley, C., Greiff, V., Heiden, J. V., Koguchi, Y., Langerak, T., Lim, T. S., Prak, E. L., Mariotti-Ferrandiz, E., Marquez, S., Meysman, P., Miho, E., Motwani, K., Nouri, N., Pavlović, M., Rubelt, F., Sandve, G. K., Schwab, N., Snapkov, I., Soto, C., Stervbo, U., Trück, J., van den Ham, H.-J., Watson, C., and Weber, C. R. The Future of Blood Testing Is the Immunome. *Frontiers in Immunology*, 12, 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.626793.

Barennes, P., Quiniou, V., Shugay, M., Egorov, E. S., Davydov, A. N., Chudakov, D. M., Uddin, I., Ismail, M., Oakes, T., Chain, B., Eugster, A., Kashofer, K., Rainer, P. P., Darko, S., Ransier, A., Douek, D. C., Klatzmann, D., and Mariotti-Ferrandiz, E. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nature Biotechnology*, 39 (2):236–245, February 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-0656-3.

Beshnova, D., Ye, J., Onabolu, O., Moon, B., Zheng, W., Fu, Y.-X., Brugarolas, J., Lea, J., and Li, B. De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Science Translational Medicine*, 12(557), August 2020. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.aaz3738.

Britanova, O. V., Putintseva, E. V., Shugay, M., Merzlyak, E. M., Turchaninova, M. A., Staroverov, D. B., Bolotin, D. A., Lukyanov, S., Bogdanova, E. A., Mamedov, I. Z., Lebedev, Y. B., and Chudakov, D. M. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *The Journal of Immunology*, 192(6):2689–2698, March 2014. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1302064.

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., and Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17(5):257–271, May 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.10.

Castro, D. C., Walker, I., and Glocker, B. Causality matters in medical imaging. *Nature Communications*, 11 (1):3673, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w.

Correa, J. D. and Bareinboim, E. From Statistical Transportability to Estimating the Effect of Stochastic Interventions. *Electronic proceedings of IJCAI 2019*, pp. 1661–1667, 2019.

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., Crawford, J. C., Clemens, E. B., Nguyen, T. H. O., Kedzierska, K., La Gruta, N. L., Bradley, P., and Thomas, P. G. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, July 2017. ISSN 1476-4687. doi: 10.1038/nature22383.

Davidsen, K., Olson, B. J., DeWitt, III, W. S., Feng, J., Harkins, E., Bradley, P., and Matsen, IV, F. A. Deep generative models for T cell receptor protein sequences. *eLife*, 8:e46935, September 2019. ISSN 2050-084X. doi: 10.7554/eLife.46935.

Dendrou, C. A., Petersen, J., Rossjohn, J., and Fugger, L. HLA variation and disease. *Nature Reviews Immunology*, 18(5):325–339, May 2018. ISSN 1474-1741. doi: 10.1038/nri.2017.143.

Dockès, J., Varoquaux, G., and Poline, J.-B. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9), September 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab055.

Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C. G., Mora, T., and Walczak, A. M. Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370 (1676):20140243, September 2015. doi: 10.1098/rstb.2014.0243.

Elhanati, Y., Sethna, Z., Callan Jr, C. G., Mora, T., and Walczak, A. M. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews*, 284(1):167–179, 2018. ISSN 1600-065X. doi: 10.1111/imr.12665.

Eliyahu, S., Sharabi, O., Elmedvi, S., Timor, R., Davidovich, A., Vigneault, F., Clouser, C., Hope, R., Nimer, A., Braun, M., Weiss, Y. Y., Polak, P., Yaari, G., and Gal-Tanamy, M. Antibody Repertoire Analysis of Hepatitis C Virus

Infections Identifies Immune Signatures Associated With Spontaneous Clearance. *Frontiers in Immunology*, 9: 3004, 2018. ISSN 1664-3224. doi: 10.3389/fimmu.2018.03004.

Emerson, R. O., DeWitt, W. S., Vignali, M., Gravley, J., Hu, J. K., Osborne, E. J., Desmarais, C., Klinger, M., Carlson, C. S., Hansen, J. A., Rieder, M., and Robins, H. S. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49(5):659–665, May 2017. ISSN 1546-1718. doi: 10.1038/ng.3822.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, April 2009. ISSN 1471-0064. doi: 10.1038/nrg2554.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits on Translational Science Proceedings*, 2020: 191–200, May 2020. ISSN 2153-4063.

Greiff, V., Bhat, P., Cook, S. C., Menzel, U., Kang, W., and Reddy, S. T. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Medicine*, 7 (1):49, May 2015. ISSN 1756-994X. doi: 10.1186/s13073-015-0169-8.

Greiff, V., Yaari, G., and Cowell, L. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, October 2020. ISSN 2452-3100. doi: 10.1016/j.coisb.2020.10.010.

Hernán, M. and Robins, J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Huang, K., Wu, L., and Yang, Y. Gut microbiota: An emerging biological diagnostic and treatment approach for gastrointestinal diseases. *JGH Open*, 5(9):973–975, 2021. ISSN 2397-9070. doi: 10.1002/jgh3.12659.

Ishigaki, K., Lagattuta, K. A., Luo, Y., James, E. A., Buckner, J. H., and Raychaudhuri, S. HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. *Nature Genetics*, pp. 1–10, March 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01032-z.

Kanduri, C., Pavlović, M., Scheffer, L., Motwani, K., Chernigovskaya, M., Greiff, V., and Sandve, G. K. Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification. *GigaScience*, 11:giac046, January 2022. ISSN 2047-217X. doi: 10.1093/gigascience/giac046. URL https://doi.org/10.1093/gigascience/giac046.

Kouw, W. M. and Loog, M. An introduction to domain adaptation and transfer learning. *arXiv:1812.11806 [cs, stat]*, December 2018.

Krishna, C., Chowell, D., Gönen, M., Elhanati, Y., and Chan, T. A. Genetic and environmental determinants of human TCR repertoire diversity. *Immunity & Ageing*, 17 (1):26, September 2020. ISSN 1742-4933. doi: 10.1186/s12979-020-00195-9.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, October 2010. ISSN 1471-0064. doi: 10.1038/nrg2825.

Liu, X., Zhang, W., Zhao, M., Fu, L., Liu, L., Wu, J., Luo, S., Wang, L., Wang, Z., Lin, L., Liu, Y., Wang, S., Yang, Y., Luo, L., Jiang, J., Wang, X., Tan, Y., Li, T., Zhu, B., Zhao, Y., Gao, X., Wan, Z., Huang, C., Fang, M., Li, Q., Peng, H., Liao, X., Chen, J., Li, F., Ling, G., Zhao, H., Luo, H., Xiang, Z., Liao, J., Liu, Y., Yin, H., Long, H., Wu, H., Yang, H., Wang, J., and Lu, Q. T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 78(8):1070–1078, August 2019. ISSN 0003-4967, 1468-2060. doi: 10.1136/annrheumdis-2019-215442.

Locke, W. J., Guanzon, D., Ma, C., Liew, Y. J., Duesing, K. R., Fung, K. Y., and Ross, J. P. DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Frontiers in Genetics*, 10:1150, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.01150.

Maros, M. E., Capper, D., Jones, D. T. W., Hovestadt, V., von Deimling, A., Pfister, S. M., Benner, A., Zucknick, M., and Sill, M. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nature Protocols*, 15(2):479–512, February 2020. ISSN 1750-2799. doi: 10.1038/s41596-019-0251-6.

Murugan, A., Mora, T., Walczak, A. M., and Callan, C. G. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1212755109.

Nestor, B., McDermott, M. B. A., Chauhan, G., Naumann, T., Hughes, M. C., Goldenberg, A., and Ghassemi,

M. Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation. *arXiv:1811.12583 [cs, stat]*, November 2018.

Nikolich-Žugich, J., Slifka, M. K., and Messaoudi, I. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2):123–132, February 2004. ISSN 1474-1741. doi: 10.1038/nri1292.

Ostmeyer, J., Christley, S., Rounds, W. H., Toby, I., Greenberg, B. M., Monson, N. L., and Cowell, L. G. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics*, 18(1):401, September 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1814-6.

Ostmeyer, J., Christley, S., Toby, I. T., and Cowell, L. G. Biophysicochemical motifs in T cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocytes and adjacent healthy tissue. *Cancer Research*, pp. canres.2292.2018, January 2019. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-18-2292.

Pavlović, M., Scheffer, L., Motwani, K., Kanduri, C., Kompova, R., Vazov, N., Waagan, K., Bernal, F. L. M., Costa, A. A., Corrie, B., Akbar, R., Al Hajj, G. S., Balaban, G., Brusko, T. M., Chernigovskaya, M., Christley, S., Cowell, L. G., Frank, R., Grytten, I., Gundersen, S., Haff, I. H., Hovig, E., Hsieh, P.-H., Klambauer, G., Kuijjer, M. L., Lund-Andersen, C., Martini, A., Minotto, T., Pensar, J., Rand, K., Riccardi, E., Robert, P. A., Rocha, A., Slabodkin, A., Snapkov, I., Sollid, L. M., Titov, D., Weber, C. R., Widrich, M., Yaari, G., Greiff, V., and Sandve, G. K. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nature Machine Intelligence*, 3(11):936–944, November 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00413-z. URL https://www.nature.com/articles/s42256-021-00413-z. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 11 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Adaptive immunity;Computational platforms and environments;Machine learning Subject_term_id: adaptive-immunity;computational-platforms-and-environments;machine-learning.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

Schaafsma, E., Fugle, C. M., Wang, X., and Cheng, C. Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy. *British Journal of Cancer*, 125(3):422–432, August 2021. ISSN 1532-1827. doi: 10.1038/s41416-021-01400-2.

Schneider-Hohendorf, T., Görlich, D., Savola, P., Kelkka, T., Mustjoki, S., Gross, C. C., Owens, G. C., Klotz, L., Dornmair, K., Wiendl, H., and Schwab, N. Sex bias in MHC I-associated shaping of the adaptive immune system. *Proceedings of the National Academy of Sciences*, 115(9):2168–2173, February 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1716146115.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2021.3058954.

Sethna, Z., Elhanati, Y., Callan, C. G., Walczak, A. M., and Mora, T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17): 2974–2981, September 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz035. URL https://academic.oup.com/bioinformatics/article/35/17/2974/5292315.

Sewell, A. K. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, September 2012. ISSN 1474-1741. doi: 10.1038/nri3279.

Shemesh, O., Polak, P., Lundin, K. E. A., Sollid, L. M., and Yaari, G. Machine Learning Analysis of Naïve B-Cell Receptor Repertoires Stratifies Celiac Disease Patients and Controls. *Frontiers in Immunology*, 12, 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.627813.

Sidhom, J.-W., Larman, H. B., Pardoll, D. M., and Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, 12(1):1605, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21879-w.

Slabodkin, A., Chernigovskaya, M., Mikocziova, I., Akbar, R., Scheffer, L., Pavlović, M., Bashour, H., Snapkov, I., Mehta, B. B., Weber, C. R., Gutierrez-Marcos, J., Sollid, L. M., Haff, I. H., Sandve, G. K., Robert, P. A., and Greiff, V. Individualized VDJ recombination predisposes the available Ig sequence space. *Genome Research*, November 2021. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.275373.121.

Storkey, A. When Training and Test Sets Are Different: Characterizing Learning Transfer. December 2008.

Subbaswamy, A. and Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, April 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz041.

Teschendorff, A. E. Avoiding common pitfalls in machine learning omic data science. *Nature Materials*, 18(5): 422–427, May 2019. ISSN 1476-4660. doi: 10.1038/s41563-018-0241-z.

Trück, J., Eugster, A., Barennes, P., Tipton, C. M., Luning Prak, E. T., Bagnara, D., Soto, C., Sherkow, J. S., Payne, A. S., Lefranc, M.-P., Farmer, A., The AIRR Community, Bostick, M., and Mariotti-Ferrandiz, E. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. *eLife*, 10:e66274, May 2021. ISSN 2050-084X. doi: 10.7554/eLife.66274.

Weber, C. R., Rubio, T., Wang, L., Zhang, W., Robert, P. A., Akbar, R., Snapkov, I., Wu, J., Kuijjer, M. L., Tarazona, S., Conesa, A., Sandve, G. K., Liu, X., Reddy, S. T., and Greiff, V. Reference-based comparison of adaptive immune receptor repertoires, January 2022. URL https://www.biorxiv.org/content/10.1101/2022.01.23.476436v1. Pages: 2022.01.23.476436 Section: New Results.

Whalen, S., Schreiber, J., Noble, W. S., and Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, pp. 1–13, November 2021. ISSN 1471-0064. doi: 10.1038/s41576-021-00434-9.

Widrich, M., Schäfl, B., Pavlović, M., Ramsauer, H., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S., and Klambauer, G. Modern Hopfield Networks and Attention for Immune Repertoire Classification. *Advances in Neural Information Processing Systems*, 33, 2020.

Yao, Y., Zia, A., Neumann, R. S., Pavlovic, M., Balaban, G., Lundin, K. E. A., Sandve, G. K., and Qiao, S.-W. T cell receptor repertoire as a potential diagnostic marker for celiac disease. *Clinical Immunology (Orlando, Fla.)*, 222:108621, January 2021. ISSN 1521-7035. doi: 10.1016/j.clim.2020.108621.

Zarnitsyna, V., Evavold, B., Schoettle, L., Blattman, J., and Antia, R. Estimating the Diversity, Completeness, and Cross-Reactivity of the T Cell Repertoire. *Frontiers in Immunology*, 4:485, 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00485.