

Estimating the Causal Effects of T Cell Receptors

Eli N. Weinstein^{*†}

Elizabeth B. Wood[†]

David M. Blei^{*‡}

October 21, 2024

Abstract

A central question in human immunology is how a patient’s repertoire of T cells impacts disease. Here, we introduce a method to infer the causal effects of T cell receptor (TCR) sequences on patient outcomes using observational TCR repertoire sequencing data and clinical outcomes data. Our approach corrects for unobserved confounders, such as a patient’s environment and life history, by using the patient’s immature, pre-selection TCR repertoire. The pre-selection repertoire can be estimated from nonproductive TCR data, which is widely available. It is generated by a randomized mutational process, V(D)J recombination, which provides a natural experiment. We show formally how to use the pre-selection repertoire to draw causal inferences, and develop a scalable neural-network estimator for our identification formula. Our method produces an estimate of the effect of interventions that add a specific TCR sequence to patient repertoires. As a demonstration, we use it to analyze the effects of TCRs on COVID-19 severity, uncovering potentially therapeutic TCRs that are (1) observed in patients, (2) bind SARS-CoV-2 antigens *in vitro* and (3) have strong positive effects on clinical outcomes.

Significance Statement. T cell receptors (TCRs) play a major role in human adaptive immunity, and increasingly form the basis for therapeutics. We propose a method to estimate the causal effects of TCRs on patient outcomes, for example, the effect of adoptive transfer of T cells with a specific TCR. Our method relies on patient TCR sequencing data, together with clinical data about patient outcomes. To correct for confounding, it uses TCR sequences with disabling mutations. The method produces a causal map from TCR sequences to patient outcomes. This work has potential future applications in designing novel TCR-T cell therapies, TCR bispecifics, T cell vaccines, and other medicines.

1 Introduction

Individual T cells’ ability to respond to disease depends crucially on their T cell receptor (TCR), a protein on the surface of the cell which recognizes antigens such as viral proteins. A patient’s collection or *repertoire* of different TCRs can shape the course of their disease, such as COVID-19, cancers, and other conditions.

^{*}Data Science Institute, Columbia University, New York, NY

[†]Jura Bio, Boston, MA

[‡]Department of Computer Science and Department of Statistics, Columbia University, New York, NY

contact: ew2760@columbia.edu, david.blei@columbia.edu

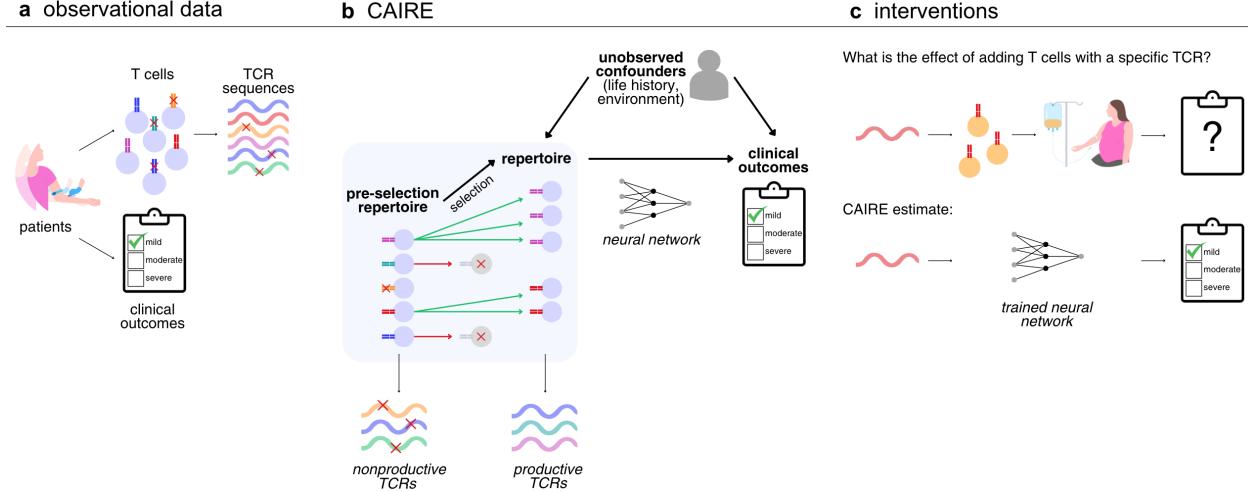


Figure 1: Estimating the causal effects of TCRs with CAIRE. (a) CAIRE uses repertoire sequencing and clinical outcomes data from patients. The sequencing data includes nonproductive TCRs. (b) CAIRE trains a neural network to estimate the effect of TCR repertoires on clinical outcomes. It uses pre-selection repertoires as an instrumental variable, to correct for unobserved confounders. The pre-selection repertoire develops into the mature repertoire through a process of antigen-dependent natural selection, in which some TCR populations expand and others die off. Productive TCR data provides information about a patient’s current repertoire; nonproductive TCR data provides information about the pre-selection repertoire. (c) CAIRE provides an estimate of the effect of giving T cells with a specific TCR to patients, e.g. via TCR-T cell therapy.

In this paper, we develop a method to estimate the causal effects of TCRs on clinical outcomes. We are not merely interested in which TCRs are associated with a disease, or which TCRs recognize disease-related antigens. Rather, our goal is to predict what would happen to patients if we *intervene* on them by adding a particular TCR to their repertoires. An accurate causal map of TCR sequences would provide a deeper understanding of human disease and directly inform the design of TCR-T cell therapies, TCR bispecifics, T cell vaccines, and other medicines [Baulu et al., 2023, Klebanoff et al., 2023].

How can we estimate the causal effects of TCRs without actually intervening on patients? Advances in high-throughput sequencing has allowed hundreds of thousands of TCRs to be sequenced from individual donors, providing a detailed picture of patients’ repertoires [Freeman et al., 2009, Robins et al., 2012]. Using this technology, several studies have collected TCR repertoire sequences from large cohorts of patients with different diseases, conditions, and outcomes [Nielsen and Boyd, 2018, Davis and Boyd, 2019, Joshi et al., 2022, Mhanna et al., 2024]. The problem we solve is how to analyze such *observational data*—observed repertoire sequences and clinical outcomes—to estimate the effects of TCR interventions.

The fundamental challenge to estimating causal effects from observational studies is *unobserved confounders*, unmeasured variables that affect both TCR repertoires, which is the treatment variable, and patient outcomes. Confounders lead to correlation without causation, so that the TCRs associated with an outcome are not necessarily responsible for that outcome. Confounding is a particular threat to repertoire studies because the human immune system is adaptive, and possesses a memory of past diseases and exposures.

For example, consider a healthcare worker who is repeatedly exposed to a variety of infectious diseases. They are likely to possess T cells protective against these previous infections, which may

be unrelated to their current disease and its symptoms. But this healthcare worker may also be more likely to have appropriate diagnosis and treatment for their disease, compared to those who do not work in healthcare. So a patient’s job can affect both their TCR repertoire and their disease outcome. Consequently, TCRs that are protective against a previous disease can be associated with clinical outcomes in a current disease, despite the fact that this relationship is not causal.

To address this confounding, we will make use of another source of data: *pre-selection repertoires*. A patient’s pre-selection repertoire is the collection of TCRs found among their T cells before those cells have encountered any antigens, that is, before the patient has been exposed to any diseases. What is important about a patient’s pre-selection repertoire is that it cannot be affected by confounders, such as job status, which only affect the current repertoire through their effects on previous diseases and exposures. Thus the pre-selection repertoire can serve as an *instrumental variable*. It is a source of randomization that affects the treatment but is unaffected by confounders and does not directly affect outcomes [Imbens and Angrist, 1994, Newey and Powell, 2003, Pearl, 2009]

Observed instrumental variables help estimate causal effects in the face of possible confounding. But how can we observe the pre-selection repertoire? In fact, there is information about patients’ pre-selection repertoire embedded in the sequencing data from their current repertoires. We extract this information to produce an observed instrumental variable, and then use this instrumental variable to estimate causal effects.

In more detail, TCRs in the pre-selection repertoire are created via *V(D)J recombination*, a mutational process in which the T cell’s genome is randomly cut up and recombined to produce a complete TCR sequence. Sometimes, V(D)J recombination produces a TCR that is not functional: it has a disabling truncation or frameshift mutation. These *nonproductive* TCRs cannot bind antigens, and so are not subject to selection. Nonproductive TCRs persist throughout the patient’s life, thus providing information about the pre-selection repertoire [Murugan et al., 2012]. Nonproductive TCRs are generally found at random in TCR sequencing data, though many analyses discard them. We use the observed nonproductive TCR data to learn about pre-selection repertoires.

Contribution. We develop CAIRE (causal adaptive immune receptor effect estimator), a method to estimate the causal effects of TCRs on clinical outcomes (Figure 1). We first show how information about patients’ pre-selection repertoire can be used to correct for unobserved confounders. Formally, we prove that the effect of interventions on TCR repertoires can be *causally identified* using a biophysical model of antigen-dependent selection. Based on this identification result, we then develop CAIRE, a practical method for using observational clinical outcomes data and TCR repertoire sequencing data to estimate the causal effect. CAIRE uses deep representation learning to scale to large, high-dimensional datasets. On semisynthetic data, we demonstrate that CAIRE can provide accurate estimates of the effects of individual TCRs on patient outcomes, even when non-causal methods are led astray. We apply CAIRE to estimate the effects of TCRs on COVID-19 severity. We use these estimates to understand how different TCRs within a patients’ repertoire contribute to their clinical outcome. We also use CAIRE’s estimates to *virtually screen* for patient TCRs that may be effective therapeutics, since they have large positive effects on clinical outcomes, and to virtually screen for antigens that may be effective vaccines, since they bind TCRs with large positive effects on clinical outcomes.

Related Work. Previous work on observational repertoire sequencing data has considered the problem of predicting a patient’s past or present disease from their repertoire sequences, and learning which TCRs are associated with that disease [Mhanna et al., 2024]. In particular, Emerson et al. [2017] pioneers diagnostic methods based on high throughput TCR sequencing, developing a

test for cytomegalovirus infection based on associations with commonly seen TCRs. [Widrich et al. \[2020\]](#) develops a neural network, transformer-based approach for predicting disease status from TCR repertoires, working in the framework of multiple instance learning. [Slabodkin et al. \[2023\]](#) employs a semi-supervised prediction approach, and shows that it can also be used to predict and generate unobserved receptor sequences associated with a disease.

Other approaches have been developed based on a variety of machine learning techniques, such as kmer frequency features, distance-based clustering, or representations derived from protein language models [[Mariotti-Ferrandiz et al., 2016](#), [Pavlović et al., 2021](#), [Zaslavsky et al., 2022](#), [Liu et al., 2024](#)]. Such methods have been applied to diseases, including COVID-19, to diagnose the disease and to identify clusters of TCRs associated with it [[Snyder et al., 2020](#), [Schultheiß et al., 2020](#), [Dannebaum et al., 2022](#), [Kockelbergh et al., 2022](#)]. As highlighted by [Pavlović et al. \[2024\]](#), however, these methods for prediction and association do not address causal questions about the effects of interventions.

Another line of work focuses on identifying receptors in an individual’s repertoire that are under strong selective pressure. In particular, [Pogorely et al. \[2018, 2019\]](#) compare the mature repertoire to an estimate of the pre-selection repertoire derived from nonproductive TCRs [[Murugan et al., 2012](#), [Marcou et al., 2018](#)]. Their purpose is to identify TCRs that have reacted against antigens the patient has encountered in the past. Here we use the pre-selection repertoire for a different purpose, to estimate causal effects.

[Pradier et al. \[2023\]](#) estimates causal effects *on* TCR repertoires, such as the effect of disease exposure, using deep generative models and causal representation learning. We consider the effects *of* repertoires, and advance methods to account for unobserved confounding.

Outside of immune repertoire studies, our work builds on causal inference methods in human genetics. The digital twin test uses parent and offspring genotyping data to infer the causal effects of genetic variants [[Bates et al., 2020](#)]. Like our method, it exploits genetic recombination as a source of randomization. It uses meiotic recombination; we use V(D)J recombination.

Our approach extends instrumental variable (IV) methods, a technique for causal inference that has been widely used in economics, epidemiology, genetics and other fields [[Imbens and Angrist, 1994](#), [Newey and Powell, 2003](#), [Davey Smith and Hemani, 2014](#), [Saengkyongam et al., 2022](#)]. Our identification results are distinct from existing IV methods, as is our approach to estimation. In particular, our results do not rely on assumptions about how the outcome variable is generated, e.g. we do not assume additive or independent noise. Instead, we use domain-specific assumptions about how the instrument affects the treatment, stemming from the biology of T cell development. Moreover, our results are not limited to binary or continuous treatments, but rather extend to complex, high-dimensional treatments, namely TCR repertoires.

Finally, this work builds on recent developments in causal inference methods for high dimensional and structured data. [Kaddour et al. \[2021\]](#) develop efficient semiparametric estimators for the effects of structured treatment data when confounding is observed; they focus on applications to graph data such as small molecules. [Hartford et al. \[2017\]](#) and [Xu et al. \[2021\]](#) extend semiparametric IV methods to high-dimensional continuous treatments, focusing especially on images. We will borrow ideas from these approaches to develop a neural-network based method for TCR effect estimation.

2 A Causal Model of Repertoires and Outcomes

Our goal is to estimate the causal effects of TCR repertoires. We take a causal graphical modeling approach, whereby we posit a causal model of our data and then use it to derive an algorithm to

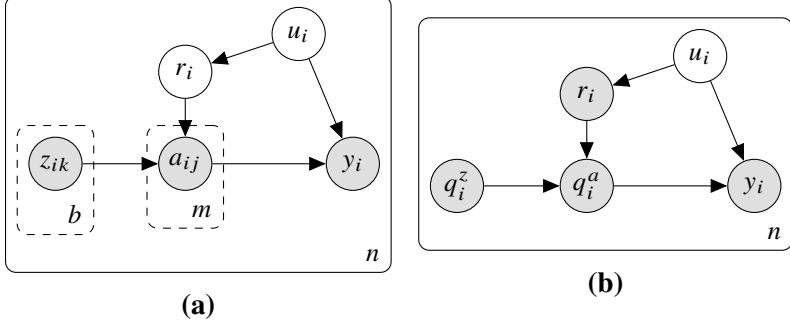


Figure 2: Causal graphs. (a) Hierarchical causal model. (b) Collapsed causal model. u_i : unobserved confounding. z_{ik} : pre-selection repertoire sequences. a_{ij} : mature repertoire sequences. y_i : patient outcomes. q_i^z : pre-selection repertoire distribution. q_i^a : repertoire distribution. r_i : relative fitness.

estimate the causal effects.

Our causal model is in Figure 2a. In this causal graph, nodes are variables, edges denote a causal relationship between them, shaded nodes are observed, and unshaded nodes are unobserved. This causal graph is hierarchical: the solid outer plate indicates replication of patients in the data; the dashed inner plates indicates replication of cells within them [Weinstein and Blei, 2024].

We describe each variable in turn, and discuss each of their roles in the causal model. Formal details on the model are in Appendix A.1.

- *Disease outcome.* The disease outcome of patient i is y_i . It is a scalar measurement, such as survival. The disease outcome is our outcome variable.
- *TCR repertoire.* The TCR repertoire of patient i is a collection of TCRs $\mathbf{a}_i = \{a_{ij}\}_{j=1}^{m_i}$. TCR a_{ij} is a sequence of amino acids, that is, a variable-length string made from a 20-letter alphabet. The variable a_{ij} falls within an inner plate of the causal graph because each patient has many T cells. There is an arrow from \mathbf{a}_i to y_i because we expect the patient’s repertoire influences their disease outcome. The repertoire is our treatment variable.
- *Unobserved confounders.* A central challenge to causal inference from observational data is unobserved confounders, variables that affect both the treatment and the outcome. For example, a patient’s environment and life history can affect both their disease outcomes and T cell repertoire. In the graph, the unobserved confounders are u_i . They connect both to the disease outcome y_i and the repertoire \mathbf{a}_i . Unobserved confounders can induce spurious (i.e., non-causal) correlation between repertoires and disease outcomes.
- *Pre-selection repertoire.* Each patient’s pre-selection repertoire is $\mathbf{z}_i = \{z_{ij}\}_{j=1}^{b_i}$. These are the receptor sequences in the patient’s *immature* T cells, i.e., those before exposure to any antigens. We will use the pre-selection repertoire as an external source of randomness that can help correct for unobserved confounding. In particular, we assume this repertoire is an *instrument*, a variable that affects the treatment but is unaffected by confounders and does not have a direct effect on the outcome [Imbens and Angrist, 1994, Newey and Powell, 2003, Pearl, 2009, Chap. 7].

Why is it a valid instrument? The pre-selection repertoire develops into the mature repertoire \mathbf{a}_i , and thus shapes its content; so we draw an arrow from \mathbf{z}_i to \mathbf{a}_i . The pre-selection repertoire is randomized, via V(D)J recombination; so we assume it is not affected by confounders, and

there is no arrow from u_i to \mathbf{z}_i . Finally, immature T cells do not respond to pathogens, and so are unlikely to directly affect a patient’s disease outcome for most diseases; thus there is no arrow from \mathbf{z}_i to y_i [Abbas et al., 2018, Chap. 8]. Taken together, these conditions imply that \mathbf{z}_i is a valid instrument. We discuss these assumptions and their limitations further in Section 6.

- *Natural selection and the fitness function.* The process of repertoire development is one of natural selection [Elhanati et al., 2014, Abbas et al., 2018]. Different TCRs in the pre-selection repertoire have different levels of fitness, and increase or decrease in number over time. The result of this selection is the mature repertoire we see today.

We explicitly account for this selection process using a latent variable $r_i(\cdot)$, which describes the relative fitness of the TCRs in a patient’s repertoire. Suppose patient i contracts influenza. Then any TCR x that recognizes the influenza virus will have high fitness relative to other TCRs, since its population is likely to expand in response to infection. For that patient, $r_i(x)$ will be large. Conversely, any TCR that recognizes patient i ’s own proteins is likely to have low fitness, since these TCRs are naturally killed off to avoid autoimmunity. For these TCRs, $r_i(x)$ will be small.

The full function r_i effectively summarizes all the selective forces that have acted on all the sequences in the patient’s repertoire over the course of their life. We assume each patient’s pre-selection repertoire \mathbf{z}_i develops into their mature repertoire \mathbf{a}_i according to a process of natural selection with relative fitness r_i .

Further, we assume that the latent fitness r_i *mediates* confounding: The confounders u_i only affect the mature repertoire \mathbf{a}_i through their effect on the fitness r_i . Our reasoning is as follows. First, a patient’s life history of antigen exposure affects the selective pressure on each TCR in their repertoire, so there must be an arrow from u_i to r_i . Second, selection shapes the mature repertoire, so there must be an arrow from r_i to \mathbf{a}_i . Finally, repertoire development is assumed to be driven by natural selection, as opposed to other biological processes [Abbas et al., 2018, Chap. 8]. So the only other way for confounders to affect the mature repertoire, besides changing TCR fitness, is by changing the initial TCR repertoire; and this we have already excluded. Hence, there is no arrow from u_i to \mathbf{a}_i . In short, we assume that the effect that any confounder – be it environment, life history, or another variable – has on the patient repertoire boils down to an effect on the relative fitness of different TCRs.

We established a causal model of TCR repertoires and disease outcomes, with the goal of estimating the causal effect of TCRs. Formally, the causal effect is described as the hypothetical result of an *intervention* on this model. Here we will estimate the effects of interventions that add TCRs with sequence a_\star to each patient’s repertoire. Concretely, one way such an intervention might be medically achieved is with TCR-T cell therapy, in which T cells engineered to possess a chosen TCR a_\star are transferred into a patient (Figure 1c) [Baulu et al., 2023].

Adapting Pearl’s do-notation, we denote the distribution of the outcome after intervention as $p(y ; \text{do}(a \sim \sigma_{a_\star, \epsilon}))$, where ϵ is a dosage parameter describing the fraction of T cells in the repertoire that have sequence a_\star after intervention [Pearl, 2009]. (We will define $\sigma_{a_\star, \epsilon}$ more formally later.) Our goal is to use data that comes from the unintervened distribution in Figure 2a to estimate $p(y ; \text{do}(a \sim \sigma_{a_\star, \epsilon}))$.

2.1 Identifying TCR Effects

The next step towards estimating the causal effect is to *causally identify* it. In causal identification, we derive a formula for the interventional distribution $p(y ; \text{do}(a \sim \sigma_{a_\star, \epsilon}))$ in terms of the distri-

bution of observable variables $p(y, \mathbf{a}, \mathbf{z})$. From this formula, we then develop estimation methods to approximate the causal effect from data (Section 3).

Causal identification usually proceeds by assuming that we see an infinite amount of data from the observational model. Figure 2a is a *hierarchical* causal model, with an inner plate. So to study identification, we consider the limit where we have data from an infinite number of TCRs within each patient ($m \rightarrow \infty$ and $b \rightarrow \infty$), as well as from an infinite number of patients ($n \rightarrow \infty$).

With this infinite data, we can effectively reconstruct the underlying *distributions* over TCR sequences within each patient's repertoire. Each patient's TCR repertoire consists of samples a_{i1}, a_{i2}, \dots from their underlying TCR distribution q_i^a ; with infinite data we effectively observe q_i^a . Likewise for pre-selection sequences $z_{ij} \sim q_i^z$, we effectively observe q_i^z . Finally, with data (q_i^z, q_i^a, y_i) from infinite patients, we effectively observe the underlying joint distribution $p(q^z, q^a, y)$. Our goal now is to write the causal effect in terms of $p(q^z, q^a, y)$.

The collapsed model and intervention. The first step of identification in a hierarchical causal model is to *collapse* the model, equating it to a flat causal model without inner plates, Figure 2b. In the collapsed model, the repertoire distributions q^a and q^z are causal variables, in place of the TCR sequences a and z . We consider the effect of the repertoire distribution q^a on y , which equates to the causal effect of the repertoire \mathbf{a} on y in the uncollapsed model. (Appendix A.3 details this equivalence.)

With the collapsed model in hand, we revisit and refine our definition of an intervention on a repertoire. First imagine we modify every T cell in every patient to have the TCR a_\star . This intervention is one where the repertoire distribution is set to $q^a = \delta_{a_\star}$, a point mass at a_\star . Of course, medically, such an intervention is challenging to achieve, and it is dangerous not to have a diverse repertoire of T cells.

Instead, we focus on more therapeutically tractable interventions, which supplement a patient's existing repertoire with a specific sequence. In particular, we consider interventions that change a patient's original repertoire distribution q_a to $q_\star^a = (1 - \epsilon)q^a + \epsilon\delta_{a_\star}$. This intervention modifies the repertoire so that a fraction ϵ of all T cells have TCR a_\star . Concretely, it might be accomplished by delivering a TCR-T cell therapeutic with TCR a_\star at a dosage ϵ . We write the intervention as $\text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})$, where $\sigma_{a_\star, \epsilon}$ is defined mathematically as the distribution over q_\star^a produced by (1) sampling q^a according to the unintervened model, $q^a \sim p(q^a | q^z, r)$, and (2) transforming q^a to $q_\star^a = (1 - \epsilon)q^a + \epsilon\delta_{a_\star}$. (Appendix A.3 equates the collapsed model intervention $\text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})$ to an intervention in the original hierarchical causal model where $a_{ij} \sim q_\star^a$, denoted $\text{do}(a \sim \sigma_{a_\star, \epsilon})$.)

An assumed model of selection. We next constrain the causal mechanism generating q^a from its parents r and q^z . This constraint, which is drawn from biological knowledge and theory, will render r to be effectively observed (Figure 2b). Without the constraint, r would be hidden, and the causal effect would not be identified [Pearl, 2009, Saengkyongam et al., 2022].

The constraint follows from the biological assumption that the mature repertoire q^a develops from the pre-selection repertoire q^z according to a process of natural selection, with fitness given by r . We mathematically express this idea with a population genetics model of evolution under natural selection [Neher and Shraiman, 2011, Bertram and Masel, 2019]. Applied to our context, the model describes a population of T cells whose genotypes are their TCR sequences [Elhanati et al., 2014, Isacchini et al., 2021].

Assumption 1 (Maturation via selection). *The causal mechanism generating q_i^a is:*

$$q_i^a(x) = f(q_i^z, r_i) = \frac{r_i(x)}{\sum_{x' \in \mathcal{X}} r_i(x') q_i^z(x')} q_i^z(x), \quad (1)$$

where \mathcal{X} is the space of sequences and $r_i(x)$ is a function $\mathcal{X} \rightarrow \mathbb{R}_+$ representing the relative fitness of sequence x in patient i .

Eq. 1 says that the fraction of TCRs with sequence x in the mature repertoire, $q_i^a(x)$, is proportional to the fraction of TCRs with that sequence in the pre-selection repertoire, $q_i^z(x)$, times the selective pressure on the sequence, $r_i(x)$. The denominator normalizes the distribution, so $\sum_{x \in \mathcal{X}} q_i^a(x) = 1$.

Assumption 1 asserts that the causal variable q^a cannot be generated from its parents q^z and r according to any arbitrary conditional distribution $p(q^a | q^z, r)$. Rather, it must be generated according to the deterministic mechanism $q^a = f(r, q^z)$. The key consequence is that the selective pressures on TCR repertoires can be reconstructed from data. Given the observed variables q^z and q^a , we can reconstruct the latent variable r as,

$$r(x) = \frac{q^a(x)/q^z(x)}{q^a(x_0)/q^z(x_0)}. \quad (2)$$

Appendix A.2 derives this fact.

Eq. 2 implies we can infer the relative fitness of each TCR by examining the likelihood ratio between the pre-selection and mature repertoire distributions. (The sequence x_0 is a reference point that can be chosen arbitrarily.) Thus, the variable r is marked as observed in Figure 2b. To be clear, Eq. 2 does not describe a causal mechanism, i.e. r is not caused by q^a and q^z . Rather, once we know q^a and q^z , we can reconstruct the value of r which led to q^a .

Complete model and identification formula. To summarize, we present the complete model.

Definition 1 (Collapsed repertoire IV model). *The collapsed repertoire IV model has the graph in Figure 2b and the following causal mechanisms:*

$$\begin{aligned} u_i &\sim p(u) \\ r_i &\sim p(r | u_i) \\ q_i^z &\sim p(q^z) \\ q_i^a(x) &= \frac{r_i(x)}{\sum_{x' \in \mathcal{X}} r_i(x') q_i^z(x')} q_i^z(x) \\ y_i &\sim p(y | q_i^a, u_i). \end{aligned} \quad (3)$$

Given q_i^z and q_i^a , TCR sequences in the pre-selection and mature repertoires are generated as $z_{ik} \sim q_i^z$ and $a_{ij} \sim q_i^a$.

We do not place any assumptions on the mechanism $p(y | q^a, u)$, so repertoires and confounders can affect the patient's outcome in any way. We do not place constraints on $p(r | u)$ either, so confounders can affect the selection pressures on TCRs in any way.

Finally, we apply the do-calculus to Figure 2b to identify the effect of intervening on patient repertoires q^a . It is identified via backdoor correction with respect to r . We can further identify the effects of interventions that add specific TCRs to patient repertoires.

Theorem 1 (TCR effects are identified). *Assume positivity: $p(q^a = q_\star^a | r) > 0$ a.s. for $r \sim p(r)$, $q_\star^a \sim \sigma_{a_\star, \epsilon}(q_\star^a | r)$. Then,*

$$p(y; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) = \int \int p(y | (1 - \epsilon)q^a + \epsilon\delta_{a_\star}, r)p(q^a, r)dq^a dr, \quad (4)$$

where $p(q^a, r)$ is derived from $p(q^a, q^z)$ via Eq. 2.

The proof is in Appendix A.3. We further discuss the positivity assumption in Appendix A.4; we discuss the result’s relationship to other hierarchical causal models in Appendix A.5.

Theorem 1 identifies the entire outcome distribution after an intervention. To summarize the effect of an intervention, we focus on the *average treatment effect*, defined as the change in the average outcome after intervention:

$$\text{ATE}(a_\star, \epsilon) \triangleq \mathbb{E}[Y; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})] - \mathbb{E}[Y; \text{do}(q_\star^a \sim \sigma_{a_\star, 0})]. \quad (5)$$

The ATE compares the average outcome after adding TCR a_\star at dosage ϵ versus the average outcome when no new TCRs are added, dosage $\epsilon = 0$.

Biological intuition. There are two complementary ways to understand the identification result in Theorem 1. One perspective is that we exploit natural variation in the pre-selection repertoire as a source of randomization, since the pre-selection repertoire is created by V(D)J recombination and unaffected by antigen exposure. Another perspective is that we exploit the imprint that a patient’s history of antigen exposures has left on their repertoire. We correct for a patient’s history of antigen exposures – and hence, confounders – by correcting for the selective forces that shaped their repertoire’s development. Following the second interpretation, we refer to our identification formula as an “antigenic history correction.” We further discuss the assumptions and limitations of the result in Section 6.

3 Estimating TCR Effects

We now show how to estimate the causal effects of TCRs from repertoire sequences and clinical data. Our method approximates the average treatment effect (Eq. 5) using the identification formula on the right hand side of Eq. 4.

For each person i we observe a bundle of data. We observe m_i TCR sequences a_{i1}, \dots, a_{im_i} from their mature repertoire. In practice these are from the CDR3 β region of the protein, which is roughly 10-20 amino acids long, and we observe about $m_i \approx 100,000$ sequences. We also observe a separate collection of \tilde{k}_i nonproductive TCR sequences $\tilde{z}_{i1}, \dots, \tilde{z}_{i\tilde{k}_i}$. Finally, we observe a clinical outcome y_i , a scalar.

The outline of our approach is as follows. First we construct a dataset of observations from the collapsed model, Figure 2b. For each patient, we estimate the mature repertoire distribution q_i^a from their productive TCR sequences and the pre-selection repertoire distribution q_i^z from their nonproductive TCR sequences. We then estimate the patient’s fitness function r_i using Eq. 2.

From these estimates, we construct a dataset $\{(\hat{q}_i^a, \hat{r}_i, y_i)\}_{i=1}^n$, with which we estimate the right side of Eq. 4. We regress y on \hat{q}^a and \hat{r} to estimate the conditional distribution of the outcome $p(y | q^a, r)$. Then we use \hat{q}^a and \hat{r} to estimate $p(q^a, r)$. Finally we combine these two estimates according to Eq. 4 to obtain the causal effect.

A significant challenge to this estimation is that the data is high dimensional. The TCR subsequences we analyze are strings of amino acids, i.e., with an alphabet of 20 characters. The identification formula in Eq. 4 involves distributions over distributions over high-dimensional discrete objects.

We use representation learning to address this challenge, employing neural networks to embed the high-dimensional sequences into a lower-dimensional space. Neural-network representations make it easier to estimate the distributions of the data; they take advantage of the assumption that similar TCR sequences have similar effects; and they allow us to estimate effects of arbitrary sequences, not only the small subset of TCRs that are commonly seen (a.k.a. *public* sequences) [Emerson et al., 2017].

Our method is called *CAIRE*: *causal adaptive immune receptor effect estimator*. In this section, we describe the statistical models underlying CAIRE, how they are fit to data, and the resulting effect estimates. Code and data is available at <https://github.com/EWeinstein/causal-tcrs>.

Estimating the repertoire distributions and fitness functions. For each patient, we need to estimate the mature repertoire distribution q_i^a and the pre-selection repertoire distribution q_i^z . To estimate q_i^a we use the empirical distribution of mature TCR sequences $\hat{q}_i^a = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta_{a_{ij}}$. To estimate q_i^z we fit a biophysical model of V(D)J recombination to the nonproductive TCR data (IGoR, [Marcou et al. \[2018\]](#)). This model extrapolates from nonproductive TCR data to an estimate of the full distribution over productive TCRs in the pre-selection repertoire \hat{q}_i^z [[Murugan et al., 2012](#)]. Note we do not have direct access to the likelihood $\hat{q}_i^z(x)$. But we can draw samples from it by sampling DNA sequences, rejecting nonproductive sequences, and translating productive DNA sequences into protein sequences.

We next estimate the fitness function r_i for each patient. Eq. 2 writes the fitness function as a density ratio between two high-dimensional distributions, q_i^a and q_i^z . Directly estimating a high-dimensional density ratio is challenging, so we use the methods of [Sugiyama et al. \[2010\]](#), [Mohamed and Lakshminarayanan \[2016\]](#) to reduce the problem to estimating a low-dimensional parameter in a classifier model.

In particular, we train a classifier to distinguish between sequences sampled from \hat{q}_i^a and those sampled from \hat{q}_i^z . Consider a candidate sequence x_{ij} and let s_{ij} denote whether it comes from the mature repertoire or the pre-selection repertoire. The model is

$$s_{ij} \sim \text{Bernoulli}(\sigma(\rho_i^\top h_r(x_{ij}; \phi) + \beta_i)), \quad (6)$$

where $\sigma(x) = 1/(1 + \exp(-x))$. Here, $h_r(x; \phi)$ is a convolutional neural network parameterized by ϕ that extracts low-dimensional features of a sequence x (Appendix E.2), while $\rho_i \in \mathbb{R}^{d_r}$ and $\beta_i \in \mathbb{R}$ are latent per-patient coefficients.

There is a one-to-one mapping between ρ_i and r_i , so we use the learned value of ρ_i as a low-dimensional representation of r_i (Appendix B.1). In particular, if the classifier is trained accurately on an equal number of samples from q_i^a and from q_i^z then Eq. 2 implies that $r_i(x) = \exp(\rho_i^\top [h_r(x; \phi) - h_r(x_0; \phi)])$. Intuitively, ρ_i describes the amount of selective pressure on each sequence feature extracted by $h_r(x; \phi)$.

Estimating the intervention. With estimates of q^a and r in hand, we now estimate the right side of Eq. 4.

First we fit a model of the outcome, $p(y | q^a, r)$:

$$y_i \sim \text{Normal}\left(\gamma_a^\top \mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] + \gamma_r^\top \rho_i + \gamma_0, \tau_y\right), \quad (7)$$

where $h_a(x; \theta)$ is another convolutional neural network, parameterized by θ . This model uses ρ_i as a representation of r_i . It uses $h_a(x; \theta)$ to extract a low-dimensional representation of sequences, and then averages individual sequence representations together to produce an overall representation of the entire repertoire, $\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)]$ [[Zaheer et al., 2017](#)]. The repertoire's representation is linear in q_i^a , which reflects the biological idea that individual TCRs act separately, i.e. TCRs do not interact with one another.

Finally we use Theorem 1 to estimate the average treatment effect (Eq. 5). We approximate the integral over $p(q^a, r)$ from the empirical distribution of \hat{q}^a and of \hat{r} , as represented by ρ .

$$\text{ATE}(a_\star, \epsilon) = \int \mathbb{E}[Y | (1 - \epsilon)q^a + \epsilon\delta_{a_\star}, r]p(q^a, r)dq^a dr - \int \mathbb{E}[Y | q^a, r]p(q^a, r)dq^a dr \quad (8)$$

$$\approx \left(\frac{1}{n} \sum_{i=1}^n \gamma_a^\top (\epsilon h_a(a_\star; \theta) + (1 - \epsilon) \mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)]) + \gamma_r^\top \rho_i + \gamma_y \right) \quad (9)$$

$$- \left(\frac{1}{n} \sum_{i=1}^n \gamma_a^\top (\mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)] + \gamma_r^\top \rho_i + \gamma_y) \right) \quad (10)$$

$$= \epsilon \gamma_a^\top \left(h_a(a_\star; \theta) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)] \right). \quad (11)$$

Summary. The method takes the following two steps. (1) Fit the parameters ρ_i , β_i and ϕ in the classifier model (Eq. 6) to estimate relative fitness r_i . (2) Fit the parameters γ_a , γ_r , γ_0 , θ and τ_y in the outcome model (Eq. 7) to predict y_i . The final result is a trained neural network $\hat{\gamma}_a^\top h_a(x; \hat{\theta})$ that predicts the effect of adding an arbitrary TCR sequence to patient repertoires (Eq. 11).

In practice, we fit the parameters of the classifier model and the outcome model simultaneously. We use stochastic optimization to scale to large datasets, drawing minibatches of patients and minibatches of repertoire sequences with each patient. We amortize inference of the per-patient latent variables ρ_i and β_i using an encoder network [Rezende et al., 2014, Kingma and Welling, 2014, Amos, 2023]. We also employ a propensity score correction, using a propensity model of the treatment q^a given the confounder r [Kaddour et al., 2021]. Further description is in Appendix B.2, and full details on architectures and training are in Appendix E.

4 Semisynthetic Data Study

We first evaluate CAIRE on semisynthetic data, where we have access to ground truth causal effects. We demonstrate that CAIRE is capable of inferring which TCRs in a patient repertoire affect the outcome, even in the presence of confounding, and even when those TCRs are rare within repertoires. We also show that the existing state of the art for repertoire classification [Widrich et al., 2020] does not provide similarly accurate causal inferences.

4.1 Semisynthetic data

To construct semisynthetic data, we extend previous methods designed for evaluating non-causal TCR repertoire classification methods [Widrich et al., 2020, Pavlović et al., 2021, Slabodkin et al., 2023]. Following these approaches, we “inject motifs” (short subsequences) into a small fraction of TCRs in a subset of patients. The presence of each motif will be associated with the outcome variable. However, unlike in previous semi-synthetic studies, only one motif will actually cause the outcome. The idea that short sequence motifs are responsible for the biological activity of TCRs has been motivated by structural studies of the binding interaction between immune receptors and antigens [Ostmeyer et al., 2019, Akbar et al., 2021, Widrich et al., 2020].

In more detail, we start with values of the pre-selection repertoire distribution q_i^z learned from a real TCR dataset [Emerson et al., 2017, Pavlović et al., 2021]. We then inject a “causal” motif, by modifying q_i^z such that a small fraction of TCRs will have a specific subsequence. Meanwhile, we set the fitness $r_i(x)$ of sequences x containing a “confounded” motif such that the fitness depends on the unobserved confounder u_i . Next, we generate q_i^a according to the selection mechanism Eq. 1.

Table 1: Method performance on semisynthetic data. Values are mean and standard error of PR AUC across 15 independent datasets.

CAIRE	Attention CAIRE	No propensity CAIRE	Uncorrected	DeepRC *
0.86 ± 0.04	0.82 ± 0.04	0.92 ± 0.03	0.56 ± 0.02	0.55 ± 0.03

We generate the outcome y_i based on the confounder u_i , and based on the fraction of sequences in the mature repertoire with the “causal” motif. In short, we design the simulation such that the presence of two motifs is associated with the outcome variable, but only one motif actually causes the outcome. Full details are in Appendix D.

4.2 Evaluation

To evaluate CAIRE, we determine how well the estimated effect $\widehat{\text{ATE}}(a_*, \epsilon)$ can discriminate sequences with the causal motif from those without it. Intuitively, we ask how well the method can recover effective therapeutic sequences—sequences that actually cause good patient outcomes—from among patient repertoires. We quantify classification performance with the area under the precision recall curve (PR-AUC) on a test set of sequences from held-out repertoires. Full details are in Appendix F.1.

4.3 Results

We compare CAIRE to several alternatives. We apply each candidate method to 15 different independently generated semisynthetic datasets, with the motifs injected into 1% of pre-selection sequences. For each method on each dataset, we use Bayesian optimization to set key hyperparameters across 10 different configurations [Balandat et al., 2020]. In CAIRE, we optimize the dimension of the latent fitness representation ρ_i and sequence representation $h_a(x; \theta)$, as well as the size of the convolutional kernel in the convolutional neural network layer of h_a . Full details on the experiments are in Appendix F.

We first demonstrate that CAIRE successfully adjusts for confounding. We compare the method to an alternative that does not account for confounding, i.e. it fixes $\rho_i = 0$ for all i . Without the confounding correction, the method is much worse at distinguishing causal TCR sequences from non-causal sequences (Table 1, CAIRE vs uncorrected).

Next we compare CAIRE to DeepRC [Widrich et al., 2020], a state-of-the-art repertoire classification method. DeepRC uses a transformer-based neural network to classify patient repertoires, e.g., learning to predict from repertoire sequences whether or not a patient has had a certain disease. The key differences between DeepRC and CAIRE are that (a) DeepRC does not correct for confounding, and (b) DeepRC computes its repertoire embedding using an attention mechanism, up-weighting certain areas of sequence space according to a learned attention score. We slightly modified data weightings in DeepRC so the method can provide valid effect estimates under the assumption of no confounding (Appendix C). We call this modified version DeepRC * . CAIRE substantially outperforms DeepRC * at causal effect estimation in the presence of confounding (Table 1).

We also compare CAIRE to a variant, “Attention CAIRE,” that corrects for confounding but uses DeepRC’s attention mechanism. Using the attention mechanism does not provide better effect estimates (Table 1).

Finally, we consider the role of the propensity model. We found a small but statistically insignificant advantage from removing the propensity model (Table 1, “no propensity CAIRE”;

permutation t-test p-value of 0.30). We include the propensity model in our method because it offers theoretical robustness guarantees, ensuring a safety net for real data.

We next considered a simulation scenario where there is in fact no confounding, i.e. u does not affect y . We evaluated CAIRE with and without its confounding correction, i.e. setting $\rho_i = 0$. We find identical performance between the two methods: both achieve an average PR-AUC of one across 10 independent simulations. So, CAIRE performs well even in the absence of confounding.

A key concern for any TCR repertoire analysis method is that it should be sensitive to small repertoire changes, since in practice the subpopulation of TCRs involved in an immune response may be small [Widrich et al., 2020, Slabodkin et al., 2023]. We lowered the motif injection rate from 1% down to 0.5% and 0.1% (Appendix F.3, Figure S4). While the performance of both CAIRE and its uncorrected version deteriorate slightly, CAIRE continues to outperform at an injection rate of 0.5%. At 0.1% both methods perform poorly, with no distinguishable difference between them (permutation t-test p-value of 0.12).

5 Application: COVID-19 Severity

We now use CAIRE to conduct an exploratory analysis of real data. We estimate the effects of TCRs on COVID-19 severity. First, we compare CAIRE’s estimates to other sources of biological information, including laboratory measurements of TCR binding. Second, we use CAIRE’s estimates to develop an overall understanding of TCR-mediated immune responses in COVID-19. Finally, we use CAIRE’s estimates to nominate therapeutic candidates, including both TCRs and vaccine antigens.

The observational dataset contains $n = 507$ COVID-19 patients from 2020 [Nolan et al., 2020, Snyder et al., 2020]. The data contains high-throughput sequencing results for the TCR β CDR3 region, with an average of 201,000 productive TCRs per patient (minimum 5,231, maximum 733,792). The study also provides clinical data about the patients’ disease outcomes. Based on this data, we constructed an overall measurement of disease severity on a three point scale, with $y = +1$ corresponding to mild disease (no hospitalization), $y = 0$ correspond to moderate disease (hospitalization), and $y = -1$ corresponding to severe disease (ICU admission and/or death). Additional details on data preprocessing and the following analysis are in Appendix I.

To evaluate predictive performance and account for uncertainty, we apply CAIRE several times to different splits of the data. We use 8-fold stratified cross validation, repeated three times with different (random) partitions of the data. We use the heldout data in each split to measure model fit. As a point estimate of the average treatment effect, we report the average of $\text{ATE}(a_*, \epsilon)$ across the ensemble of 24 effect estimates. As a measure of uncertainty, we estimate the probability \tilde{p} that the sign of the point estimate is incorrect, based on a Gaussian fit to the ensemble of effect estimates (Appendix I.4) [Lakshminarayanan et al., 2017, Wilson and Izmailov, 2020].

We present our findings in detail in the following sections. The main results are as follows:

- TCRs have diverse effects: among sequences found in patient repertoires, a substantial number have modest positive effects on patient outcomes, but some have strong positive effects and some have negative effects.
- Causal effects in patients are distinct from laboratory binding measurements: the causal effect of a TCR sequence is related to its binding properties but does not perfectly correlate.
- Patients have heterogenous repertoires: all patients contain a mix of TCRs with different causal effects, although some patients with severe COVID-19 have an especially large number of TCRs with negative effects.

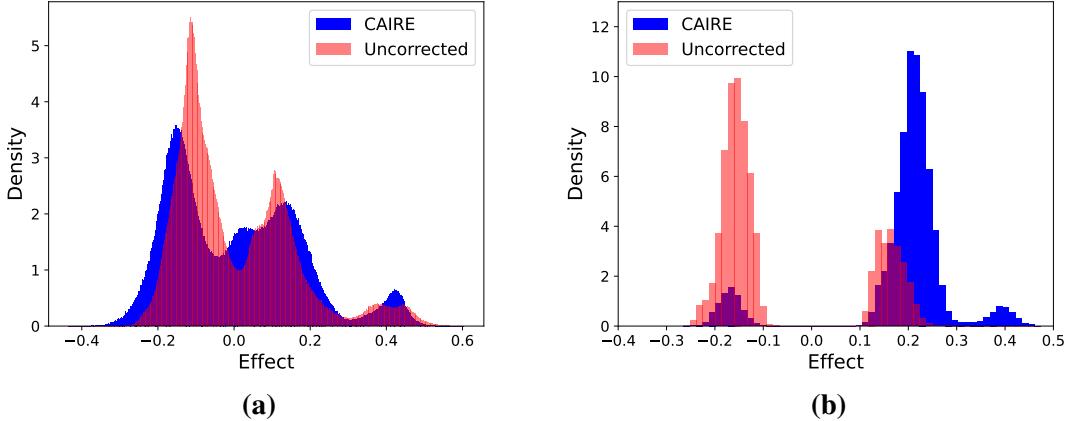


Figure 3: Distribution of TCR effects in held-out patient repertoires. (a) Effect distribution across all repertoire sequences. (b) Effect distribution across sequences with significant effects ($\tilde{p} < 0.05$)

- Patient repertoires contain promising therapeutic candidates: we uncover TCR sequences that are (1) observed in patients, (2) possess *in vitro* experimental evidence for binding SARS-CoV-2 epitopes, and (3) possess observational clinical evidence for efficacy, based on CAIRE.

5.1 Effect estimates

We use CAIRE to estimate the effect $\text{ATE}(a_\star; \epsilon)$ of interventions that use each of 11 million TCR sequences. These sequences come from the repertoires of 71 held-out COVID patients (Appendix I.3). We set the dosage parameter ϵ to 0.1, based on the dosage of existing TCR-based therapies (Appendix I.5). Note that the effect at other ϵ values can also be obtained by linearly rescaling (Eq. 11). We find 170,000 sequences (1.6% of the total) with significant effects on outcomes ($\tilde{p} < 0.05$). Among these, there is a small population of sequences with a strong positive effect (≈ 0.4 on the three-point severity scale), a larger population with a weak positive effect (≈ 0.2), and a small population with a weak negative effect (≈ -0.15) (Figure 3).

We compare CAIRE’s estimates of causal effects to those of a method without the antigenic history correction (“Uncorrected”). We observe a substantial shift in the distribution of effects: the uncorrected method predicts that, among sequences with a significant effect, 70% have a negative effect on patient outcomes, while CAIRE predicts only 10% (Figure 3b). This suggests that many sequences are statistically associated with severe disease not because they cause severe disease, but rather due to confounding. For example, this may be due to greater viral exposure in patients with severe COVID-19, leading to greater positive selection of virus-responsive TCRs, regardless of whether those TCRs are actually causing worse clinical outcomes. Overall, CAIRE’s results are consistent with the established idea that T cell-mediated immunity is broadly important in combatting SARS-CoV-2, while the uncorrected method’s estimates are not [Moss, 2022].

CAIRE finds substantial confounding: its confounder term (i.e., $\gamma_r^\top \rho_i$) explains roughly twice as much of the variance in the outcome as the treatment term (Table 3, Figure S7). However, in this dataset, the confounding captured by CAIRE does not appear to reflect easy-to-measure demographic variables. We do not find evidence that the confounder representation ρ_i is associated with age, gender or ethnicity (Appendix I.7).

The causal effect estimates provided by CAIRE are largely robust to different model architectures. We first compared CAIRE to the variant that incorporates attention. “Attention CAIRE” achieves a similar fit to the data (Table 3, Figure S7), and its effect estimates show a strong correlation with those of CAIRE (Pearson R of 0.95, p value for non-correlation below floating point precision; Figure S8a). We also compared CAIRE to a much simpler method, that does not use any non-linearities, though it still includes convolutions (Appendix I.2). This “Non-Neural CAIRE” achieves a somewhat worse fit to the data (Table 3, Figure S7), and its effect estimates are moderately correlated with CAIRE’s (Pearson R of 0.69, p value below precision; Figure S8b).

5.2 Comparison to laboratory experiments

We compare the causal effects of TCRs, estimated by CAIRE, to laboratory measurements of TCR binding with SARS-CoV-2 antigens. While *in vitro* binding does not necessarily imply that a TCR will have a strong effect on patient outcomes, it is reasonable to expect the causal effect of a TCR to be related to its ability to bind viral proteins. We use data collected from a high-throughput experiment (MIRA, multiplex identification of T-cell receptor antigen specificity) which identifies patient TCRs that bind different antigens from across the SARS-CoV-2 genome [Nolan et al., 2020, Snyder et al., 2020]. The assay provides the sequences of TCRs that bind each antigen (32,770 sequences total).

We evaluate how well CAIRE’s effect estimates predict *in vitro* TCR binding. This evaluation is not trivial because the assay does not provide sequences found to *not* bind viral antigens. However, we do have repertoire sequencing data collected from the patients whose T cells were used in the binding assay. That is, we have access to sequence data without binding labels, but drawn from the same underlying distribution of sequences as used in the binding experiments. In Appendix H we show how this unlabeled data can be used to evaluate a method’s ability to discriminate binders from non-binders. In particular, we show that the ROC AUC of a classifier evaluated on discriminating known binders from unlabelled sequences (MIRA hits from repertoire sequences) is a conservative estimate of the classifier’s AUC when evaluated on discriminating binders from non-binders.

We find that that CAIRE’s effect estimates can discriminate class I MHC binders with an ROC AUC of 0.563 ± 0.003 (standard error; Appendix I.8). Restricting the analysis just to sequences with significant effect estimates ($\tilde{p} < 0.05$), the AUC is 0.604 ± 0.027 . By contrast, the uncorrected method achieves a slightly lower AUC on all sequences (0.556 ± 0.003) and a still lower AUC on sequences with high-confidence effects (0.534 ± 0.020). These results suggest that (a) CAIRE’s effect estimates are somewhat consistent with, though not identical to, *in vitro* experimental binding data, and (b) the antigenic history correction can lead to effect estimates that correspond more closely to experimental data.

5.3 Balanced immune responses

We next try to better understand how patients’ repertoires, taken as a whole, drive disease. Our results, presented below, support the following conceptual model for TCR-mediated immune responses in COVID-19. First, in general, patient repertoires provide TCRs capable of fighting the virus. However, this success comes at a cost: repertoires also carry a burden of TCRs with negative effects, likely because these TCRs produce overly extreme immune responses. Some burden of negative effect TCRs is present even in individuals with mild disease, but in a subset of patients it is large, and helps drive severe disease. It is the balance of the two factors – TCRs that drive safe immune responses and TCRs that drive damaging immune responses – that determines how the repertoire as a whole affects patient outcomes.

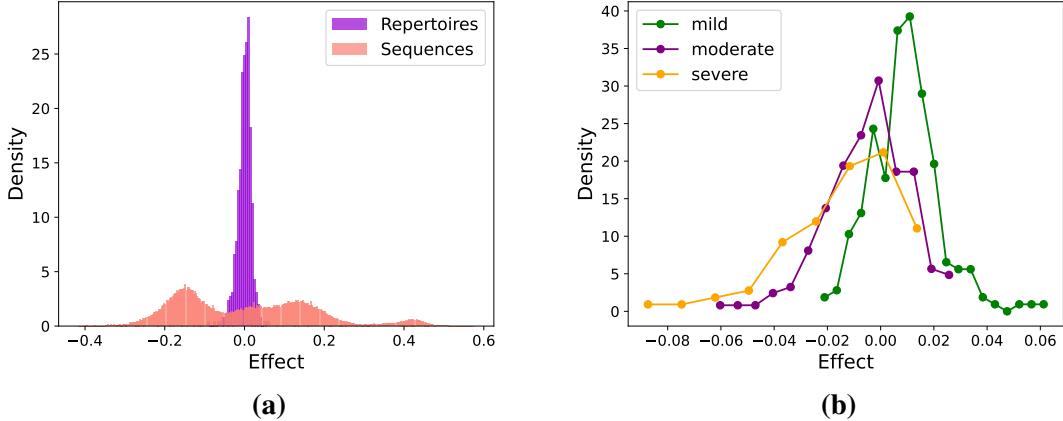


Figure 4: Effect heterogeneity within and between patients (a) The distribution of average effects across repertoires (purple) and the average distribution of effects within repertoires (red). More precisely, each bar of the purple histogram covering interval \mathcal{I} is an estimate of $\mathbb{P}_{Q^a \sim p(q^a)}[\mathbb{E}_{A \sim Q^a}[\text{ATE}(A; 0.1)] \in \mathcal{I}]$. Each bar of the red histogram is an estimate of $\mathbb{E}_{Q^a \sim p(q^a)}[\mathbb{P}_{A \sim Q^a}[\text{ATE}(A; 0.1) \in \mathcal{I}]]$. (b) Distribution of average effects across repertoires, from patients with different outcomes. Each point at interval \mathcal{I} is an estimate of $\mathbb{P}_{Q^a \sim p(q^a|y)}[\mathbb{E}_{A \sim Q^a}[\text{ATE}(A; 0.1)] \in \mathcal{I}]$ for an outcome $y \in \{-1, 0, +1\}$.

We applied CAIRE to estimate the effects, $\text{ATE}(a_\star, \epsilon = 0.1)$, of repertoire sequences drawn from each of the $n = 507$ COVID-19 patients in the training data. We compared variability of effects *between* patients to variability *within* patients (Figure 4a). We find that there is much more variation in the effects of individual TCRs within patients than there is variation in the average effect of repertoires between patients ($\sqrt{\mathbb{V}_{p(q^a)}[\mathbb{E}_{Q^a}[\text{ATE}(A, 0.1)]]} = 0.018$ versus $\sqrt{\mathbb{E}_{p(q^a)}[\mathbb{V}_{Q^a}[\text{ATE}(A, 0.1)]]} = 0.12$, where \mathbb{V} denotes variance). Indeed, most patients contain a mix of TCRs with large positive and large negative effects, but these balance each other out to create a more modest average effect for the repertoire overall (Figures 4b and S10a). (Note the effect of a patient’s entire repertoire, $\mathbb{E}[Y ; \text{do}(q_\star^a = q_i^a)] - \mathbb{E}[Y] = \mathbb{E}_{q_i^a}[\text{ATE}(A, 1)] = 10 \times \mathbb{E}_{q_i^a}[\text{ATE}(A, 0.1)]$ has an estimated average absolute value of $\mathbb{E}_{p(q^a)}|\mathbb{E}_{Q^a}[\text{ATE}(A, 1)]| \approx 0.13$.)

Some TCRs are predicted to have a negative effect on disease outcomes, suggesting they do more harm than good in responding to viral infection. Although the antigenic history correction in CAIRE reduces the size of this population, it does not eliminate it altogether. Biologically, one possible explanation for negative causal effects is that these TCRs drive an overactive immune response. This is consistent with the clinical observation that some patients with severe COVID-19 suffer from symptoms such as severe hyperinflammation and cytokine storm, leading to conditions such as acute respiratory distress syndrome or multisystem inflammatory syndrome [Cheng et al., 2020, Lucas et al., 2020, Kalfaoglu et al., 2021, Mobasheri et al., 2022, Moss, 2022]. COVID-19 patients also have increased risk of autoimmune disease post-infection [Sharma and Bayry, 2023].

We find that all patients have some sequences with negative effects, but a subpopulation of patients with severe or moderate COVID-19 possess a greatly expanded set of TCRs with strong negative effects (Figures 4b and S10c). Moreover, TCRs with negative causal effects are more likely to bind SARS-CoV-2 antigens, as compared to TCRs with positive effects (Table 4 column 1, Figure S9). This supports the hypothesis that TCRs can have a negative effect because they induce too wide and overactive an immune response [Stone et al., 2015, Shakiba et al., 2022].

In summary, our results suggest that the role of a patient’s TCR repertoire in COVID-19

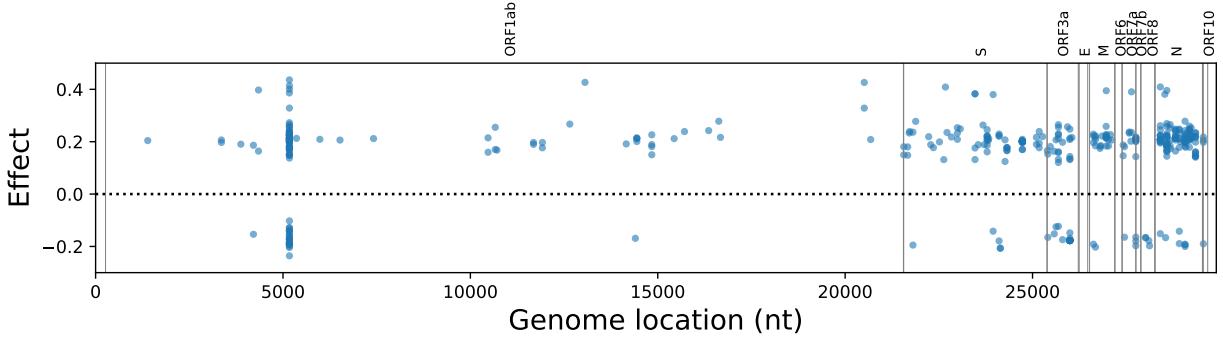


Figure 5: Distribution of TCR effects across the SARS-CoV-2 genome. x-axis: antigen location in the SARS-CoV-2 genome, indexed by nucleotide. y-axis: estimated causal effect of TCRs that bind that antigen. Each dot represents an individual TCR with a significant effect ($\tilde{p} < 0.05$) that was found to bind an epitope encoded at the given location in the SARS-CoV-2 genome.

depends on the balance between two populations of TCRs: those that drive safe and effective immune responses versus those that drive damaging immune responses. Each individual patient possesses a different repertoire and hence a different balance, shaping their individual outcome.

5.4 Implications for therapeutics and vaccines

We next consider the implications of CAIRE’s causal estimates for therapeutics. We first consider therapies based on adding TCRs to a patient’s repertoire, such as cell therapies or TCR bispecifics [Papayanni et al., 2021, Papadopoulou et al., 2023, Verhagen et al., 2021]. Applying CAIRE to binders identified in MIRA assays, we find 17 candidate therapeutic sequences with strong positive effects ($\text{ATE}(a_*) > 0.3$, $\tilde{p} < 0.05$; Table 5). These candidates are (1) observed in human patients (2) possess *in vitro* experimental support for their binding activity (from MIRA), and (3) possess observational clinical support for therapeutic efficacy (from CAIRE).

More broadly, our results caution against relying too much on *in vitro* assays of on-target binding when developing candidate therapeutics: 60% of binders found in MIRA assays are estimated to have *negative* clinical effects, while another 37% are predicted to have only weak clinical effects ($0 < \text{ATE}(a_*) < 0.3$). We discuss the implications of our results for therapeutics that *subtract* TCRs from a patient’s repertoire in Appendix I.9 [Moisini et al., 2008, Norville and Wood, 2023].

We next consider the implications of CAIRE’s estimates for vaccine development. A central question in vaccine design is how to choose an antigen that induces a strong, beneficial immune response. To address this question, we examine the effects of TCRs that bind each SARS-CoV-2 antigen studied in the MIRA binding assay. We find many antigens that bind a diversity of TCRs, including TCRs with significant positive and significant negative effects (Figure 5). However, a subset of antigens were enriched for TCRs with significant positive effects (Table 6; binomial test, Benjamini-Hochberg adjusted p-value below 0.05). These may be promising vaccine candidates. Among these antigens are one spike protein epitope, as well as three nucleocapsid epitopes, including NP_{105–113}. This last epitope was identified in experimental studies as a candidate for protection against severe COVID-19 [Peng et al., 2020, 2022]. Overall, these results support the hypothesis that, besides the spike protein, the nucleocapsid protein may be a good candidate for improved T-cell vaccines [Moss, 2022].

6 Discussion

We propose a new method for causal inference from observational TCR repertoire data. We use V(D)J recombination as a source of randomness. We show that a standard biophysical model of T cell development implies causal identification. We then develop an estimation strategy using nonproductive TCR data. The method is general purpose in that it only requires TCR sequencing data and clinical outcomes. It is also scalable, running on data with ~100 million sequences.

Assumptions and limitations. As with all observational causal inference techniques, our method comes with assumptions and limitations. Our identification approach assumes that the process of V(D)J recombination is unaffected by confounders. But we cannot entirely exclude the possibility of hidden confounding: the biological mechanisms that lead to variation among individuals' pre-selection repertoires are poorly understood, and there may be some genetic or environmental factors that affect both the process of V(D)J recombination and patient outcomes [Slabodkin et al., 2021]. Another limitation is that CAIRE may control for instruments, as some factors that affect TCR fitness are likely unrelated to patient outcomes; correcting for these factors risks worse effect estimates [Bhattacharya and Vogt, 2007]. Moreover, our model of T cell repertoire development accounts only for selection, and ignores processes such as genetic drift that may also play an important role in development [Horns et al., 2019, Koraichi et al., 2023]. Also, our approach assumes that the process of V(D)J recombination is stable over time, such that nonproductive TCRs provide reliable evidence about the past immature T cells which gave rise to current mature cells. Finally, the performance of CAIRE depends on the performance of the upstream computational tools it relies on, in particular the methods used to reconstruct TCR clonotype sequences from raw sequencing data, and to reconstruct the pre-selection repertoire from nonproductive sequences.

Future work. To advance precision medicine applications, it will be important to develop conditional treatment effect estimates rather than only study average treatment effects. A key challenge in particular is to estimate the effect of TCRs conditional on a patient's HLA type, as HLAs directly affect TCR-antigen binding. Finally, although we have focused on T cell receptors, CAIRE may also be applied to B cell receptors. An additional challenge here is that somatic hypermutation plays an important role in B cell receptors, complicating the assumption that repertoire development proceeds by selection alone. But effect estimates for B cell receptors have the potential to inform antibody-based medicines, a major class of therapeutics.

7 Acknowledgments

We wish to thank Alan Amin, Andrei Slabodkin, and Mattia Gollub for useful discussions.

References

- Abul K Abbas, Andrew H Lichtman, and Shiv Pillai. *Cellular and Molecular Immunology*. Elsevier, ninth edition, 2018.
- Rahmad Akbar, Philippe A Robert, Milena Pavlović, Jeliazko R Jeliazkov, Igor Snapkov, Andrei Slabodkin, Cédric R Weber, Lonneke Scheffer, Enkelejda Miho, Ingrid Hobæk Haff, Dag Trygve Tryslew Haug, Fridtjof Lund-Johansen, Yana Safonova, Geir K Sandve, and Victor Greiff. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep.*, 34(11):108856, March 2021.

Brandon Amos. Tutorial on amortized optimization. *Found. Trends. Mach. Learn.*, 16(5):592–732, June 2023.

Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, February 2016.

Maximilian Balandat, Brian Karrer, Daniel R Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Neural Information Processing Systems*, 2020.

Stephen Bates, Matteo Sesia, Chiara Sabatti, and Emmanuel Candès. Causal inference in genetic trio studies. *Proc. Natl. Acad. Sci. U. S. A.*, 117(39):24117–24126, September 2020.

Estelle Baulu, Célia Gardet, Nicolas Chuvin, and Stéphane Depil. TCR-engineered T cell therapy in solid tumors: State of the art and perspectives. *Sci. Adv.*, 9(7), February 2023.

Jason Bertram and Joanna Masel. Density-dependent selection and the limits of relative fitness. *Theor. Popul. Biol.*, 129:81–92, October 2019.

Jay Bhattacharya and William B Vogt. Do instrumental variables belong in propensity scores? *NBER Technical Working Paper Series*, 2007.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20(28):1–6, 2019.

Mary Hongying Cheng, She Zhang, Rebecca A Porritt, Magali Noval Rivas, Lisa Paschold, Edith Willscher, Mascha Binder, Moshe Arditi, and Ivet Bahar. Superantigenic character of an insert unique to SARS-CoV-2 spike supported by skewed TCR repertoire in patients with hyperinflammation. *Proc. Natl. Acad. Sci. U. S. A.*, 117(41):25254–25262, October 2020.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 10093–10100, 2020.

Corinna Cortes and M Mohri. Confidence intervals for the area under the ROC curve. *Neural Information Processing Systems*, 2004.

Richard Dannebaum, Phillip Suwalski, Hosseinali Asgharian, Gracie Du Zhipei, Hai Lin, Januay Weiner, Manuel Holtgrewe, Charlotte Thibeault, Melina Müller, Xiaomin Wang, Zehra Karadeniz, Jacopo Saccomanno, Jan-Moritz Doehn, Ralf-Harto Hübner, Bernd Hinzenmann, Anja Blüher, Sandra Siemann, Dilduz Telman, Norbert Suttorp, Martin Witzenrath, Stefan Hippensiel, Carsten Skurk, Wolfgang Poller, Leif E Sander, Dieter Beule, Florian Kurth, Toumy Guettouche, Ulf Landmesser, Jan Berka, Khai Luong, Pa-COVID Study Group, Florian Rubelt, and Bettina Heidecker. Highly multiplexed immune repertoire sequencing links multiple lymphocyte classes with severity of response to COVID-19. *EClinicalMedicine*, 48:101438, June 2022.

George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.*, 23(R1):R89–98, September 2014.

Mark M Davis and Scott D Boyd. Recent progress in the analysis of $\alpha\beta$ T cell and B cell receptor repertoires. *Curr. Opin. Immunol.*, 59:109–114, August 2019.

A P Dawid. Influence diagrams for causal modelling and inference. *Int. Stat. Rev.*, 70(2):161–189, 2002.

Yuval Elhanati, Anand Murugan, Curtis G Callan, Jr, Thierry Mora, and Aleksandra M Walczak. Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci. U. S. A.*, 111(27):9875–9880, July 2014.

Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, Mark Rieder, and Harlan S Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, 49(5):659–665, May 2017.

David Freedman and Persi Diaconis. On the histogram as a density estimator: L₂ theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, December 1981.

J Douglas Freeman, René L Warren, John R Webb, Brad H Nelson, and Robert A Holt. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.*, 19(10):1817–1824, October 2009.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, 2017.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, 89(22):10915–10919, November 1992.

Felix Horns, Christopher Vollmers, Cornelia L Dekker, and Stephen R Quake. Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. *Proc. Natl. Acad. Sci. U. S. A.*, 116(4):1261–1266, January 2019.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

Giulio Isacchini, Aleksandra M Walczak, Thierry Mora, and Armita Nourmohammad. Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proc. Natl. Acad. Sci. U. S. A.*, 118(14), April 2021.

Kroopa Joshi, Martina Milighetti, and Benjamin M Chain. Application of T cell receptor (TCR) repertoire analysis for the advancement of cancer immunotherapy. *Curr. Opin. Immunol.*, 74:1–8, February 2022.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. In *Neural Information Processing Systems*, 2021.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellemudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. A study of BFLOAT16 for deep learning training. *arXiv*, 2019.

Bahire Kalfaoglu, José Almeida-Santos, Chanidapa Adele Tye, Yorifumi Satou, and Masahiro Ono. T-cell dysregulation in COVID-19. *Biochem. Biophys. Res. Commun.*, 538:204–210, January 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, June 2019.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-Normalizing neural networks. In *Neural Information Processing Systems*, 2017.

Christopher A Klebanoff, Smita S Chandran, Brian M Baker, Sergio A Quezada, and Antoni Ribas. T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome. *Nat. Rev. Drug Discov.*, pages 1–22, October 2023.

Mark Klinger, Katherine Kong, Martin Moorhead, Li Weng, Jianbiao Zheng, and Malek Faham. Combining next-generation sequencing and immune assays: a novel method for identification of antigen-specific T cells. *PLoS One*, 8(9):e74231, September 2013.

Mark Klinger, Francois Pepin, Jen Wilkins, Thomas Asbury, Tobias Wittkop, Jianbiao Zheng, Martin Moorhead, and Malek Faham. Multiplex identification of Antigen-Specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One*, 10(10): e0141561, October 2015.

Hannah Kockelbergh, Shelley Evans, Tong Deng, Ella Clyne, Anna Kyriakidou, Andreas Economou, Kim Ngan Luu Hoang, Stephen Woodmansey, Andrew Foers, Anna Fowler, and Elizabeth J Soilleux. Utility of bulk T-Cell receptor repertoire sequencing analysis in understanding immune responses to COVID-19. *Diagnostics (Basel)*, 12(5), May 2022.

Meriem Bensouda Koraichi, Silvia Ferri, Aleksandra M Walczak, and Thierry Mora. Inferring the T cell repertoire dynamics of healthy individuals. *Proceedings of the National Academy of Sciences*, 120(4):e2207516120, 2023.

B Lakshminarayanan, A Pritzel, and C Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Neural Information Processing Systems*, 2017.

Meiling Liu, Yang Liu, Li Hsu, and Qianchuan He. TCRpred: incorporating T-cell receptor repertoire for clinical outcome prediction. *Front. Genet.*, 15, March 2024.

Carolina Lucas, Patrick Wong, Jon Klein, Tiago B R Castro, Julio Silva, Maria Sundaram, Mallory K Ellingson, Tianyang Mao, Ji Eun Oh, Benjamin Israelow, Takehiro Takahashi, Maria Tokuyama, Peiwen Lu, Arvind Venkataraman, Annsea Park, Subhasis Mohanty, Haimei Wang, Anne L Wyllie, Chantal B F Vogels, Rebecca Earnest, Sarah Lapidus, Isabel M Ott, Adam J Moore,

M Catherine Muenker, John B Fournier, Melissa Campbell, Camila D Odio, Arnau Casanovas-Massana, Yale IMPACT Team, Roy Herbst, Albert C Shaw, Ruslan Medzhitov, Wade L Schulz, Nathan D Grubaugh, Charles Dela Cruz, Shelli Farhadian, Albert I Ko, Saad B Omer, and Akiko Iwasaki. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature*, July 2020.

Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-París. How many TCR clonotypes does a body maintain? *J. Theor. Biol.*, 389:214–224, January 2016.

Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, 9(1):561, February 2018.

Encarnita Mariotti-Ferrandiz, Hang-Phuong Pham, Sophie Dulauroy, Olivier Gorgetto, David Klatzmann, Pierre-Andre Cazenave, Sylviane Pied, and Adrien Six. A TCR β repertoire signature can predict experimental cerebral malaria. *PLoS One*, 11(2):e0147871, February 2016.

Vanessa Mhanna, Habib Bashour, Khang Lê Quý, Pierre Barennes, Puneet Rawat, Victor Greiff, and Encarnita Mariotti-Ferrandiz. Adaptive immune receptor repertoire analysis. *Nature Reviews Methods Primers*, 4(1):1–25, January 2024.

Leila Mobasher, Mohammad Hossein Nasirpour, Elham Masoumi, Afsaneh Foolady Azarnaminy, Mozhdeh Jafari, and Seyed-Alireza Esmaeili. SARS-CoV-2 triggering autoimmune diseases. *Cytokine*, 154:155873, June 2022.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models, 2016.

Ioana Moisini, Phuong Nguyen, Lars Fugger, and Terrence L Geiger. Redirecting therapeutic T cells against myelin-specific T lymphocytes using a humanized myelin basic protein-HLA-DR2-zeta chimeric receptor. *J. Immunol.*, 180(5):3601–3611, March 2008.

Paul Moss. The T cell immune response against SARS-CoV-2. *Nat. Immunol.*, 23(2):186–193, February 2022.

Iván Díaz Muñoz and Mark van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, June 2012.

Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan, Jr. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U. S. A.*, 109(40):16161–16166, October 2012.

Richard A Neher and Boris I Shraiman. Statistical genetics and evolution of quantitative traits. *Rev. Mod. Phys.*, 83(4):1283–1300, November 2011.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, September 2003.

Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemøller, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, 12(5):1007–1017, May 2003.

Sandra C A Nielsen and Scott D Boyd. Human adaptive immune receptor repertoire analysis-past, present, and future. *Immunol. Rev.*, 284(1):9–23, July 2018.

Sean Nolan, Marissa Vignali, Mark Klinger, Jennifer N Dines, Ian M Kaplan, Emily Svejnoha, Tracy Craft, Katie Boland, Mitch Pesesky, Rachel M Gittelman, Thomas M Snyder, Christopher J Gooley, Simona Semprini, Claudio Cerchione, Massimiliano Mazza, Ottavia M Delmonte, Kerry Dobbs, Gonzalo Carreño-Tarragona, Santiago Barrio, Vittorio Sambri, Giovanni Martinelli, Jason D Goldman, James R Heath, Luigi D Notarangelo, Jonathan M Carlson, Joaquin Martinez-Lopez, and Harlan S Robins. A large-scale database of T-cell receptor beta (TCR β) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq*, August 2020.

Julie Norville and Elizabeth Wood. Major histocompatibility complex-based chimeric receptors and uses thereof for treating autoimmune diseases. *United States Patent*, 11826385, November 2023.

Jared Ostmeyer, Scott Christley, Inimary T Toby, and Lindsay G Cowell. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res.*, 79(7):1671–1680, April 2019.

Anastasia Papadopoulou, George Karavalakis, Efthymia Papadopoulou, Aliki Xochelli, Zoi Bousiou, Anastasios Vogiatzoglou, Penelope-Georgia Papayanni, Aphrodite Georgakopoulou, Maria Giannaki, Fani Stavridou, Ioanna Vallianou, Maria Kammenou, Evangelia Varsamoudi, Vasiliki Papadimitriou, Chrysavgi Giannaki, Maria Sileli, Zoi Stergiouda, Garyfallia Stefanou, Georgia Kourlaba, George Gounelas, Maria Triantafyllidou, Eleni Siotou, Antonia Karaglani, Eleni Zotou, Georgia Chatzika, Anna Boukla, Apostolia Papalexandri, Maria-Georgia Koutra, Dimitra Apostolou, Georgia Pitsiou, Petros Morfesis, Michalis Doumas, Theodoros Karampatakis, Nikolaos Kapravelos, Militsa Bitzani, Maria Theodorakopoulou, Eva Serasli, Grigorios Georgolopoulos, Ioanna Sakellari, Asimina Fylaktou, Stavros Tryfon, Achilles Anagnostopoulos, and Evangelia Yannaki. SARS-CoV-2-specific T cell therapy for severe COVID-19: a randomized phase 1/2 trial. *Nat. Med.*, 29(8):2019–2029, August 2023.

Penelope-Georgia Papayanni, Dimitrios Chasiotis, Kiriakos Koukoulias, Aphrodite Georgakopoulou, Anastasia Iatrou, Eleni Gavriliaki, Chrysavgi Giannaki, Militsa Bitzani, Eleni Geka, Polychronis Tasioudis, Diamantis Chloros, Asimina Fylaktou, Ioannis Kioumis, Maria Triantafyllidou, Sotiria Dimou-Besikli, Georgios Karavalakis, Afroditi K Boutou, Eleni Siotou, Achilles Anagnostopoulos, Anastasia Papadopoulou, and Evangelia Yannaki. Vaccinated and convalescent donor-derived severe acute respiratory syndrome coronavirus 2-specific T cells as adoptive immunotherapy for high-risk coronavirus disease 2019 patients. *Clin. Infect. Dis.*, 73(11):2073–2082, December 2021.

Milena Pavlović, Lonneke Scheffer, Keshav Motwani, Chakravarthi Kanduri, Radmila Kompovala, Nikolay Vazov, Knut Waagan, Fabian L M Bernal, Alexandre Almeida Costa, Brian Corrie, Rahmad Akbar, Ghadi S Al Hajj, Gabriel Balaban, Todd M Brusko, Maria Chernigovskaya, Scott Christley, Lindsay G Cowell, Robert Frank, Ivar Grytten, Sveinung Gundersen, Ingrid Hobæk Haff, Eivind Hovig, Ping-Han Hsieh, Günter Klambauer, Marieke L Kuijjer, Christin Lund-Andersen, Antonio Martini, Thomas Minotto, Johan Pensar, Knut Rand, Enrico Riccardi, Philippe A Robert, Artur Rocha, Andrei Slabodkin, Igor Snapkov, Ludvig M Sollid, Dmytro Titov, Cédric R Weber, Michael Widrich, Gur Yaari, Victor Greiff, and Geir Kjetil Sandve. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nature Machine Intelligence*, 3(11):936–944, November 2021.

Milena Pavlović, Ghadi S Al Hajj, Chakravarthi Kanduri, Johan Pensar, Mollie E Wood, Ludvig M Sollid, Victor Greiff, and Geir K Sandve. Improving generalization of machine learning-identified biomarkers using causal modelling with examples from immune receptor diagnostics. *Nature Machine Intelligence*, 6(1):15–24, January 2024.

Judea Pearl. *Causality*. Cambridge University Press, September 2009.

Yanchun Peng, Alexander J Mentzer, Guihai Liu, Xuan Yao, Zixi Yin, Danning Dong, Wanwisa Dejnirattisai, Timothy Rostron, Piyada Supasa, Chang Liu, César López-Camacho, Jose Slon-Campos, Yuguang Zhao, David I Stuart, Guido C Paesen, Jonathan M Grimes, Alfred A Antson, Oliver W Bayfield, Dorothy E D P Hawkins, De-Sheng Ker, Beibei Wang, Lance Turtle, Krishan-thi Subramaniam, Paul Thomson, Ping Zhang, Christina Dold, Jeremy Ratcliff, Peter Simmonds, Thushan de Silva, Paul Sopp, Dannielle Wellington, Ushani Rajapaksa, Yi-Ling Chen, Mariolina Salio, Giorgio Napolitani, Wayne Paes, Persephone Borrow, Benedikt M Kessler, Jeremy W Fry, Nikolai F Schwabe, Malcolm G Semple, J Kenneth Baillie, Shona C Moore, Peter J M Openshaw, M Azim Ansari, Susanna Dunachie, Eleanor Barnes, John Frater, Georgina Kerr, Philip Goulder, Teresa Lockett, Robert Levin, Yonghong Zhang, Ronghua Jing, Ling-Pei Ho, Oxford Immunology Network Covid-19 Response T cell Consortium, ISARIC4C Investigators, Richard J Cornall, Christopher P Conlon, Paul Klenerman, Gavin R Screaton, Juthathip Mongkolsapaya, Andrew McMichael, Julian C Knight, Graham Ogg, and Tao Dong. Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. *Nat. Immunol.*, 21(11):1336–1345, November 2020.

Yanchun Peng, Suet Ling Felce, Danning Dong, Frank Penkava, Alexander J Mentzer, Xuan Yao, Guihai Liu, Zixi Yin, Ji-Li Chen, Yongxu Lu, Dannielle Wellington, Peter A C Wing, Delaney C C Dominey-Foy, Chen Jin, Wenbo Wang, Megat Abd Hamid, Ricardo A Fernandes, Beibei Wang, Anastasia Fries, Xiaodong Zhuang, Neil Ashley, Timothy Rostron, Craig Waugh, Paul Sopp, Philip Hublitz, Ryan Beveridge, Tiong Kit Tan, Christina Dold, Andrew J Kwok, Charlotte Rich-Griffin, Wanwisa Dejnirattisa, Chang Liu, Prathiba Kurupati, Isar Nassiri, Robert A Watson, Orion Tong, Chelsea A Taylor, Piyush Kumar Sharma, Bo Sun, Fabiola Curion, Santiago Revale, Lucy C Garner, Kathrin Jansen, Ricardo C Ferreira, Moustafa Attar, Jeremy W Fry, Rebecca A Russell, COMBAT Consortium, Hans J Stauss, William James, Alain Townsend, Ling-Pei Ho, Paul Klenerman, Juthathip Mongkolsapaya, Gavin R Screaton, Calliope Dendrou, Stephen N Sansom, Rachael Bashford-Rogers, Benny Chain, Geoffrey L Smith, Jane A McKeating, Benjamin P Fairfax, Paul Bowness, Andrew J McMichael, Graham Ogg, Julian C Knight, and Tao Dong. An immunodominant NP105-113-B*07:02 cytotoxic T cell response controls viral replication and is associated with less severe COVID-19 disease. *Nat. Immunol.*, 23(1):50–61, January 2022.

Scott Pesme, Aymeric Dieuleveut, and Nicolas Flammariou. On convergence-diagnostic based step sizes for stochastic gradient descent. *International Conference on Machine Learning*, 2020.

Mikhail V Pogorelyy, Anastasia A Minervina, Dmitriy M Chudakov, Ilgar Z Mamedov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Method for identification of condition-associated public antigen receptor sequences. *Elife*, 7, March 2018.

Mikhail V Pogorelyy, Anastasia A Minervina, Mikhail Shugay, Dmitriy M Chudakov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.*, 17(6):e3000314, June 2019.

Melanie F Pradier, Niranjani Prasad, Paidamoyo Chapfuwa, Sahra Ghalebikesabi, Maximilian Ilse, Steven Woodhouse, Rebecca Elyanow, Javier Zazo, Javier Gonzalez Hernandez, Julia Greissl, and Edward Meeds. AIRIVA: A deep generative model of adaptive immune repertoires. In *Machine Learning for Healthcare Conference*, 2023.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Harlan Robins, Cindy Desmarais, Jessica Matthis, Robert Livingston, Jessica Andriesen, Helena Reijonen, Christopher Carlson, Gerold Nepom, Cassian Yee, and Karen Cerosaletti. Ultra-sensitive detection of rare T cell clones. *J. Immunol. Methods*, 375(1-2):14–19, January 2012.

Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, and Jonas Peters. Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning*, 2022.

Christoph Schultheiß, Lisa Paschold, Donjete Simnica, Malte Mohme, Edith Willscher, Lisa von Wenserski, Rebekka Scholz, Imke Wieters, Christine Dahlke, Eva Tolosa, Daniel G Sedding, Sandra Ciesek, Marylyn Addo, and Mascha Binder. Next-generation sequencing of T and B cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. *Immunity*, 53(2):442–455.e4, August 2020.

Mojdeh Shakiba, Paul Zumbo, Gabriel Espinosa-Carrasco, Laura Menocal, Friederike Dündar, Sandra E Carson, Emmanuel M Bruno, Francisco J Sanchez-Rivera, Scott W Lowe, Steven Camara, Richard P Koche, Vincent P Reuter, Nicholas D Socci, Benjamin Whitlock, Fella Tamzalit, Morgan Huse, Matthew D Hellmann, Daniel K Wells, Nadine A Defranoux, Doron Betel, Mary Philip, and Andrea Schietinger. TCR signal strength defines distinct mechanisms of T cell dysfunction and cancer evasion. *J. Exp. Med.*, 219(2), February 2022.

Chetan Sharma and Jagadeesh Bayry. High risk of autoimmune diseases after COVID-19. *Nat. Rev. Rheumatol.*, 19(7):399–400, July 2023.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Association for the Advancement of Artificial Intelligence*, 2006.

Andrei Slabodkin, Maria Chernigovskaya, Ivana Mikocziava, Rahmad Akbar, Lonneke Scheffer, Milena Pavlović, Habib Bashour, Igor Snapkov, Brij Bhushan Mehta, Cédric R Weber, Jose Gutierrez-Marcos, Ludvig M Sollid, Ingrid Hobæk Haff, Geir Kjetil Sandve, Philippe A Robert, and Victor Greiff. Individualized VDJ recombination predisposes the available ig sequence space. *Genome Res.*, 31(12):2209–2224, November 2021.

Andrei Slabodkin, Ludvig M Sollid, Geir Kjetil Sandve, Philippe Auguste Robert, and Victor Greiff. Weakly supervised identification and generation of adaptive immune receptor sequences associated with immune disease status. September 2023.

Thomas M Snyder, Rachel M Gittelman, Mark Klinger, Damon H May, Edward J Osborne, Ruth Taniguchi, H Jabran Zahid, Ian M Kaplan, Jennifer N Dines, Matthew T Noakes, Ravi Pandya, Xiaoyu Chen, Summer Elasady, Emily Svejnoha, Peter Ebert, Mitchell W Pesesky, Patricia De Almeida, Hope O'Donnell, Quinn DeGottardi, Gladys Keitany, Jennifer Lu, Allen Vong, Rebecca Elyanow, Paul Fields, Julia Greissl, Lance Baldo, Simona Semprini, Claudio Cerchione, Fabio Nicolini, Massimiliano Mazza, Ottavia M Delmonte, Kerry Dobbs, Rocio Laguna-Goya, Gonzalo Carreño-Tarragona, Santiago Barrio, Luisa Imberti, Alessandra Sottini, Eugenia Quiros-Roldan, Camillo Rossi, Andrea Biondi, Laura Rachele Bettini, Mariella D'Angio, Paolo Bonfanti, Miranda F Tompkins, Camille Alba, Clifton Dalgard, Vittorio Sambri, Giovanni Martinelli, Jason D Goldman, James R Heath, Helen C Su, Luigi D Notarangelo, Estela Paz-Artal, Joaquin Martinez-Lopez, Jonathan M Carlson, and Harlan S Robins. Magnitude and dynamics of the T-Cell response to SARS-CoV-2 infection at both individual and population levels. *medRxiv*, September 2020.

Jennifer D Stone, Daniel T Harris, and David M Kranz. TCR affinity for p/MHC formed by tumor antigens that are self-proteins: impact on efficacy and toxicity. *Curr. Opin. Immunol.*, 33:16–22, April 2015.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio estimation: A comprehensive review. *RIMS Kokyuroku*, 1703:10–31, 2010.

Johan Verhagen, Edith D van der Meijden, Vanessa Lang, Andreas E Kremer, Simon Völkl, Andreas Mackensen, Michael Aigner, and Anita N Kremer. Human CD4+ T cells specific for dominant epitopes of SARS-CoV-2 spike and nucleocapsid proteins with therapeutic potential. *Clin. Exp. Immunol.*, 205(3):363–378, September 2021.

Eli N Weinstein and David M Blei. Hierarchical causal models. *arXiv*, 2024.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, and Others. Modern Hopfield networks and attention for immune repertoire classification. *Neural Information Processing Systems*, 2020.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. 2020.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Neural Information Processing Systems*, 2017.

Maxim E Zaslavsky, Nikhil Ram-Mohan, Joel M Guthridge, Joan T Merrill, Jason D Goldman, Ji-Yeon Lee, Krishna M Roskin, Charlotte Cunningham-Rundles, M Anthony Moody, Barton F Haynes, Benjamin A Pinsky, James R Heath, Judith A James, Samuel Yang, Catherine A Blish, Robert Tibshirani, Anshul Kundaje, and Scott D Boyd. Disease diagnostics using machine learning of immune receptors. April 2022.

Supplementary Material

A Details on Identification

A.1 Hierarchical causal model

In this section we define the causal model and interventions formally.

Definition S2 (Repertoire IV model). *The repertoire IV model has the graph shown in Figure 2a and hierarchical causal graphical model (HCGM) equations are,*

$$\begin{aligned} u_i &\sim p(u) \\ r_i &\sim p(r \mid u_i) \\ q_i^z &\sim p(q^z) & z_{ik} &\sim q_i^z \\ q_i^a &\sim p(q^a \mid \{z_{ik}\}_{k=1}^b, r_i) & a_{ij} &\sim q_i^a \\ y_i &\sim p(y \mid \{a_{ij}\}_{j=1}^m, u_i), \end{aligned} \tag{12}$$

for $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$ and $k \in \{1, \dots, b\}$.

We use the notation $\{\cdot\}_{j=1}^m$ to indicate that the causal mechanisms $p(y \mid \{a_{ij}\}_{j=1}^m, u_i)$ and $p(q^a \mid \{z_{ik}\}_{k=1}^b, u_i)$ cannot depend on the ordering of the TCRs, just on the unordered set of sequences.¹ This reflects the idea that there is no natural ordering of a patient's T cells.

We consider interventions where a single TCR sequence a_\star is added to a patient's repertoire. Formally, this corresponds to a stochastic or *soft* intervention, since there is still some randomness in the treatment variable after the intervention, coming from the rest of the repertoire [Chap. 4 Pearl, 2009, Dawid, 2002, Muñoz and van der Laan, 2012, Correa and Bareinboim, 2020].

Definition S3 (Intervention by adding a TCR). *After intervention, each patient's repertoire distribution is generated according to the conditional distribution $\sigma_{a_\star, \epsilon}(q_\star^a \mid r)$, defined as*

$$\begin{aligned} q^a &\sim p(q^a \mid r) \\ q_\star^a &= (1 - \epsilon)q^a + \epsilon\delta_{a_\star}, \end{aligned} \tag{13}$$

where $p(q^a \mid r)$ is the causal mechanism in the un-intervened model (Definition S2). In other words, $\sigma_{a_\star, \epsilon}$ is the pushforward of $p(q^a \mid r)$ through the function $(1 - \epsilon)q^a + \epsilon\delta_{a_\star}$. After intervention, each patient's repertoire is generated as

$$q_i^a \sim \sigma_{a_\star, \epsilon}(q_\star^a \mid r_i) \quad a_{ij} \sim q_i^a$$

for $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$

A.2 Fitness and natural selection

In this section, we review the mathematical model of evolution under natural selection used in Eq. 1, and the derivation of Eq. 2 [Neher and Shraiman, 2011, Bertram and Masel, 2019].

Fitness is a measure of how much an organism with a particular genotype reproduces. In our setting, the organism is a T cell and the genotype is their TCR sequence. Fitness $g_i(x)$ is the size

¹Technically, $\{a_{ij}\}_{j=1}^m$ is a multiset, as some sequences can be identical and in this case the number of repeated sequences is still relevant.

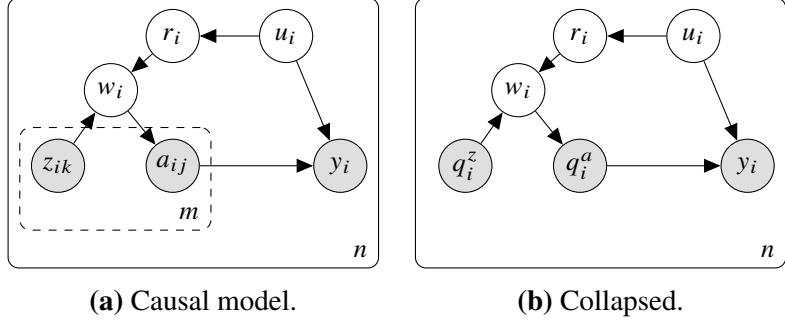


Figure S1: The repertoire IV causal model, rewritten.

of the population of mature cells with TCR gene sequence x divided by the size of the population of immature cells with sequence x . If s_i is the total number of cells in the pre-selection repertoire, then $s_i q_i^z(x)$ is the total number of cells with sequence x , and $g_i(x) s_i q_i^z(x)$ is the total number of cells with sequence x in the post-selection repertoire. So, the distribution of sequences $q_i^a(x)$ in the post-selection repertoire must be proportional to $g_i(x) q_i^z(x)$ up to a normalizing constant, i.e. $q_i^a(x) \propto g_i(x) q_i^z(x)$.

To describe fitness in a way that is agnostic to the normalizing constant, we employ not *absolute* fitness $g_i(x)$ but instead *relative* fitness $r_i(x)$, which is the ratio of the fitness of sequence x to the fitness of an arbitrarily chosen reference sequence, x_0 , i.e. $r_i(x) = g_i(x)/g_i(x_0)$. From $q_i^a(x) \propto g_i(x) q_i^z(x)$ and $\sum_x q_i^a(x) = 1$ we have,

$$q_i^a(x) = \frac{g_i(x) s_i}{\sum_{x' \in \mathcal{X}} g_i(x') s_i q_i^z(x')} q_i^z(x),$$

and Eq. 1 follows.

Since $r_i(x_0) = 1$, we also have from plugging x_0 into Eq. 1 that $\sum_{x' \in \mathcal{X}} r_i(x') q_i^z(x') = q_i^z(x_0)/q_i^a(x_0)$. So,

$$q_i^a(x) = \frac{r_i(x)}{q_i^z(x_0)/q_i^a(x_0)} q_i^z(x).$$

Solving for r_i yields Eq. 2.

A.3 Proof of identification (Theorem 1)

In this section we prove Theorem 1. The first step is to rewrite the hierarchical causal model in the format of Def. 3 in [Weinstein and Blei \[2024\]](#). Next we introduce a regularity assumption which ensures the infinite repertoire limit exists, and derive the collapsed model. Then we apply do-calculus to identify the effect of hard interventions on q^a . Finally we apply σ -calculus to identify the effect of the intervention of interest.

Rewritten HCM We rewrite the causal model (Definition S2) so that it explicitly meets Def. 3 of [Weinstein and Blei \[2024\]](#), and we can apply the identification techniques developed in that paper. We assume $b = m$, i.e. the number of pre-selection repertoire sequences matches the number of mature repertoire sequences. Since the identification result takes the number of sequences to infinity, this is a minor restriction; it can be relaxed and identification will still hold.

Definition S4 (Rewritten repertoire IV model). *The repertoire IV model (Definition S2) is equivalent to an HCGM with graph in Figure S1a and equations,*

$$\begin{aligned}
u_i &\sim p(u) \\
q_i^z &\sim p(q^z) \quad z_{ik} \sim q_i^z \\
w_i &\sim p(q^a \mid \{z_{ik}\}_{k=1}^m, u_i) \\
q_i^a &\sim \delta_{w_i} \quad a_{ij} \sim q_i^a \\
y_i &\sim p(y \mid \{a_{ij}\}_{j=1}^m, u_i).
\end{aligned} \tag{14}$$

for $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$.

Here w_i is an unobserved *interferer* variable. Intuitively, there is interference between the pre-selection and mature repertoires because multiple mature T cells can descend from the same immature T cell, resulting in multiple copies of the same TCR in the mature repertoire. This model is equivalent to the original repertoire IV model, as we have simply rewritten the mechanism for a in terms of an intermediate variable w . We can see the model meets Def. 3 in [Weinstein and Blei \[2024\]](#).

Infinite subunit limit We study identification in the limit of infinite subunits, i.e. sequences. We assume the limit exists.

Assumption S2 (Mechanism convergence). *Assume the mechanisms of the model in Definition S4 converge in Kullback-Leibler divergence, such that the model meets the conditions of Theorem 1 of Weinstein and Blei [2024].*

The exact conditions are defined and discussed in detail in [Weinstein and Blei \[2024\]](#).

Theorem 1 of [Weinstein and Blei \[2024\]](#) implies that the hierarchical causal model converges in the limit $m \rightarrow \infty$ to a collapsed model with graph in Figure S1b and equations,

$$\begin{aligned}
u_i &\sim p(u) \\
q_i^z &\sim p(q^z) \\
r_i &\sim p(r \mid u_i) \\
w_i &\sim p(q^a \mid q_i^z, r_i) \\
q_i^a &\sim \delta_{w_i} \\
y_i &\sim p(y \mid q_i^a, u_i).
\end{aligned} \tag{15}$$

Marginalizing out w_i and imposing Assumption 1 on the mechanism $p(q^a \mid q^z, r)$, we recover the collapsed repertoire IV model (Definition 1).

Hard interventions on repertoires We first identify the effect of an intervention that draws the repertoire sequences for each patient from a fixed distribution, $p(y ; \text{do}(a \sim q_\star^a))$. This is equivalent to a hard intervention on the repertoire distribution $p(y ; \text{do}(q^a = q_\star^a))$.

Theorem S1 (Repertoire effects are identified). *Assume the repertoire IV model (Definition S2) satisfies Assumption S2 and Assumption 1. Further assume positivity, i.e. $p(q_\star^a \mid r) > 0$ a.s. for $r \sim p(r)$. Then, the causal effect of hard interventions on the repertoire is identified from $p(y, q^a, q^z)$ as,*

$$p(y ; \text{do}(q^a = q_\star^a)) = \int p(y \mid q^a = q_\star^a, r)p(r)dr, \tag{16}$$

where $p(r)$ is derived from $p(q^a, q^z)$ via Eq. 2.

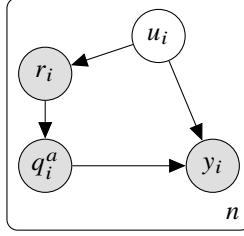


Figure S2: Marginalized collapsed repertoire IV model.

Proof. By Thm. 1 of [Weinstein and Blei \[2024\]](#), it suffices to identify the effect in the collapsed model, Definition 1, since the effect in the hierarchical causal model will be equivalent to the effect in the collapsed model in the limit $m \rightarrow \infty$. Now, marginalize out the variable q_i^z from the collapsed model, so that the mechanism generating q^a from its parents (that is, r) is stochastic (Figure S2). Eq. 16 follows by an application of do-calculus to Figure S2, under the positivity conditions of [Shpitser and Pearl \[2006\]](#) (Assumption 3 in [Weinstein and Blei \[2024\]](#)). \square

Identification proof We now prove the main identification result. Interventions that add a TCR correspond to soft interventions on the repertoire distribution. So, we apply σ -calculus, an extension of do-calculus to soft interventions [[Correa and Bareinboim, 2020](#), Thm. 1].

Proof. Starting from the collapsed and marginalized model in Figure S2, we decompose the effect as,

$$\begin{aligned} p(y; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) \\ = \int \int p(y | r, q^a; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) p(q^a | r; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) p(r; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) dr dq^a. \end{aligned}$$

From rule 2 of σ -calculus, we have $p(y | r, q^a; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) = p(y | r, q^a)$. From rule 3 of σ -calculus, we have $p(r; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) = p(r)$. So,

$$p(y; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) = \int p(y | q^a, r) p(q^a | r; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) p(r) dq^a dr.$$

The conditional distribution $p(q^a | r; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon}))$ is specified by the intervention, Eq. 13. Plugging in Eq. 13 yields the identifying equation,

$$p(y; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) = \int p(y | (1 - \epsilon)q^a + \epsilon\delta_{a_\star}, r) p(q^a, r) dq^a dr.$$

To be able to compute the integral on the right hand side, it must be possible to identify $p(y | (1 - \epsilon)q^a + \epsilon\delta_{a_\star}, r)$ for all values of q^a, r with $p(q^a, r) > 0$ (see e.g. Appx. H in [Weinstein and Blei \[2024\]](#) for further discussion). This is guaranteed by the stated positivity condition, which ensures that the intervention has non-zero probability under the observational distribution: $p((1 - \epsilon)q^a + \epsilon\delta_{a_\star}, r) > 0$ for $q^a, r \sim p(q^a, r)$ a.s.. \square

A.4 Positivity

For causal identification to hold, the intervention must satisfy positivity, i.e. it must naturally occur under the observational distribution. In particular, the intervention must have positive probability

regardless of the value of the adjustment variable, r . Here we show that it is indeed possible to satisfy the positivity condition. We focus on the positivity condition used for hard repertoire interventions (Theorem S1), as it is easier to interpret, but the same reasoning extends to interventions that add a TCR (Theorem 1). For simplicity, here we take the set of all sequences \mathcal{X} to be finite.

Proposition S1. *Assume the set of all sequences \mathcal{X} is finite. Assume $p(q^z)$ has full support over distributions on sequences, i.e. $p(q^z) > 0$ for all $q^z \in \mathcal{P}(\mathcal{X})$. Assume that with probability 1 under $p(r)$, $r(x) > 0$ for all $x \in \mathcal{X}$. Then, $p(q^a = q_\star^a | r) > 0$ for $r \sim p(r)$ a.s..*

Proof. Since r must be strictly positive, the selection process can always be run in reverse with $1/r$ in place of r , giving

$$\tilde{q}^z(x) = f(r^{-1}, q_\star^a) = \frac{r^{-1}(x)}{\sum_{x'} r^{-1}(x') q_\star^a(x')} q_\star^a(x).$$

This distribution $\tilde{q}^z(x)$ generates the intervention, $q_\star^a = f(r, \tilde{q}^z)$. Since \tilde{q}^z is a valid distribution, it has positive probability, $p(\tilde{q}^z) > 0$. The conclusion follows. \square

It is useful to contrast this result with an alternative model, where instead of the relative fitness mechanism (Eq. 1) we assume a mechanism of the form $q_i^a(x) = r_i(x)q_i^z(x)$, without the normalizer in Eq. 1. In this case, $r_i(x) = q_i^a(x)/q_i^z(x)$, the likelihood ratio. Now, however, $\tilde{q}^z(x) = r_i^{-1}(x)q_\star^a(x)$ is not necessarily a valid distribution, as it may not be normalized. As a result, positivity would be difficult to guarantee.

A.5 Relationship to HCM IV

We briefly discuss the relationship of the repertoire IV model presented here to the hierarchical IV model described in [Weinstein and Blei \[2024\]](#). Though both methods use hierarchical causal models with a subunit-level instrument and treatment variables, they rely on different types of datasets and different identification assumptions.

The HCM IV model in [Weinstein and Blei \[2024\]](#) assumes that it is possible to observe both the treatment and the instrument within each subunit. This implies that it is possible to estimate the joint distribution over the treatment and instrument within each unit, $q_i(a, z)$. The identification result depends crucially on knowledge of this joint, as it involves a backdoor correction on the conditional $q_i(a | z)$.

In the repertoire IV model presented here, however, we do not have direct access to the joint distribution over pre-selection and mature sequences, since we cannot observe those immature T cells which die off. Instead, we use domain knowledge to constrain the relationship between the instrument and the treatment (Assumption 1). As a result, in the identification formula of Theorem S1, the relative fitness r_i occupies essentially the same role as the conditional $q_i(a | z)$ in the HCM IV identification formula.

B Further Description of Estimation

B.1 Selection estimate

Here we describe our estimation strategy for relative fitness in more depth.

To construct the statistical model in Eq. 6, we assign each of the mature repertoire sequences a_{i1}, \dots, a_{im_i} the label $s_{ij} = 1$, and then draw an equal number of samples z_{i1}, \dots, z_{im_i} from \hat{q}_i^z and

assign them the label $s_{ij} = 0$. Let $\{(x_{ij}, s_{ij})\}_{j=1}^{2m_i} = \{(a_{ij}, 1)\}_{j=1}^{m_i} \cup \{(z_{ij}, 0)\}_{j=1}^{m_i}$ denote the pooled dataset. We now construct the model in Eq. 6, restated here:

$$s_{ij} \sim \text{Bernoulli}(\sigma(\rho_i^\top h_r(x_{ij}; \phi) + \beta_i)),$$

for $j \in \{1, \dots, 2m_i\}$ and $i \in \{1, \dots, n\}$. Note for fixed ϕ , this is a hierarchical logistic regression model, with ρ_i and β_i the per-patient regression coefficients and offset respectively. If h_r is unrestricted, then this model can describe any conditional distribution $p(s | x)$, since s is binary.

We now relate the representation ρ_i to the relative fitness r_i . Let $q_i(s, x)$ be the true joint distribution over sequences and labels for patient i , and assume the regression model accurately matches $q_i(s | x)$. Applying Bayes' rule,

$$\frac{q_i^a(x)}{q_i^z(x)} = \frac{q_i(x | s = 1)}{q_i(x | s = 0)} = \frac{q_i(s = 1 | x)}{q_i(s = 0 | x)} \frac{q_i(s = 0)}{q_i(s = 1)} = \frac{\sigma(\rho_i^\top h_r(x; \phi) + \beta_i)}{1 - \sigma(\rho_i^\top h_r(x; \phi) + \beta_i)} = \exp(\rho_i^\top h_r(x; \phi) + \beta_i), \quad (17)$$

where we have used the fact that $q_i(s = 0) = q_i(s = 1) = 0.5$ by construction. Plugging this into Eq. 2, we find that,

$$r_i(x) = \exp(\rho_i^\top [h_r(x; \phi) - h_r(x_0; \phi)]). \quad (18)$$

Note β_i does not appear. This equation implies a one-to-one relationship between ρ_i and r_i , under the minor regularity condition that each feature of $h_r(x; \phi)$ is unique, i.e. there is no $k \neq k' \in \{1, \dots, d_r\}$ such that $h_r(x; \phi)_k = h_r(x; \phi)_{k'}$ for all $x \in \mathcal{X}$. So, we use ρ_i as our representation of r_i . Intuitively, ρ_i is a vector describing the amount of selection in patient i on the sequence features $h_r(x; \phi)$.

B.2 Regularization and training

One possible concern is that our estimates may be poor if the outcome model is misspecified, or converges slowly to the truth. To help address these concerns, we implement a propensity score correction. Such corrections can help improve the efficiency and robustness of causal estimates, by reducing estimators' sensitivity to nuisance parameters, i.e. those parameters that do not enter into the effect itself. We use a method based on the Robinson decomposition and Neyman orthogonality, “structured intervention networks” [Kaddour et al., 2021]. This method allows for high-dimensional, structured treatments. The basic idea is to build a propensity model of the treatment’s representation, $\mathbb{E}_{q_i^a}[h_a(A; \theta)]$, and use this to adjust the regularization of the outcome model. This approach, based on semiparametric theory, can offer many of the benefits of standard propensity score methods developed for low-dimensional treatments, such as reduced bias from regularization, improved robustness to model misspecification, and fast convergence guarantees.

With this propensity correction in place, our complete model is,

$$\begin{aligned} s_{ij} &\sim \text{Bernoulli}(\sigma(\rho_i^\top h_r(x_{ij}; \phi) + \beta_i)) \\ y_i &\sim \text{Normal}\left(\gamma_a^\top (\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] - W \cdot \rho_i - B) + \gamma_r^\top \rho_i + \gamma_y, \tau_y\right), \end{aligned} \quad (19)$$

where $\{(x_{ij}, s_{ij})\}_{j=1}^{2m_i} = \{(a_{ij}, 1)\}_{j=1}^{m_i} \cup \{(z_{ij}, 1)\}_{j=1}^{m_i}$. The parameters W and B are learned from a propensity model which predicts the treatment representation $\mathbb{E}_{q_i^a}[h_a(A; \theta)]$ from the confounder representation ρ_i ,

$$\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] \sim \text{Normal}(W \cdot \rho_i + B, \tau_e). \quad (20)$$

The intuition behind the propensity-corrected term $\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] - W \cdot \rho_i - B$ in Eq. 19 is that it describes just those aspects of the treatment that are not explained by the confounder ρ_i . Note that

the propensity model describes the learned representation $\mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)]$ rather than actual data; it is an auxiliary tool for de-biasing the main model (Eq. 19), not part of the main model’s generative description of the data.

To train the entire model on datasets with large numbers of patients and TCRs, we use a stochastic inference procedure. We draw minibatches of patients, and for each patient we draw a minibatch of mature TCR sequences and a minibatch of simulated pre-selection repertoire sequences. We use these minibatches to form an approximation of the log likelihood of the entire dataset. The latent variables ρ_i and β_i are local, per-patient variables, so we use amortized inference with an encoder network. We place priors on all parameters and perform maximum *a posteriori* inference, learning point estimates of each. For optimization we employ AMSGrad, an extension of Adam with improved convergence guarantees [Reddi et al., 2018]. We train the main model (Eq. 19) and the propensity model (Eq. 20) in tandem, following the procedure of Kaddour et al. [2021]. In particular, we alternate between (1) updating the parameters of the propensity model, W , B and τ_e , based on the likelihood of the propensity model alone, and (2) updating the rest of the parameters based on the likelihood of the main model alone. The propensity model updates are performed less often than the main model updates (every 10 steps).

C DeepRC and DeepRC *

In this section we describe the DeepRC repertoire classification model proposed by Widrich et al. [2020], and our modified version, DeepRC * , which can be used to estimate causal effects under the assumption of no confounding.

We first briefly introduce some additional definitions. Let $\{\tilde{a}_{ij} : j \in \tilde{m}_i\}$ denote the set of unique sequences in the data, and let c_{ij} be the number of observed counts of sequence \tilde{a}_{ij} . The empirical distribution of the data can then be written in terms of c as $\hat{q}_i^a = \frac{1}{m_i} \sum_{j=1}^{\tilde{m}_i} \delta_{a_{ij}} = \frac{1}{m_i} \sum_{j=1}^{\tilde{m}_i} c_{ij} \tilde{a}_{ij}$, where $\tilde{m}_i = |\{\tilde{a}_{ij} : j \in \tilde{m}_i\}|$ is the number of unique sequences.

DeepRC The DeepRC architecture employs an attention mechanism. To define it, we introduce the function,

$$g(a; \theta, \eta) \equiv \exp(\eta_2^\top \tilde{h}(h_a(a; \theta); \eta_1) / \sqrt{\tilde{d}}) \quad (21)$$

Here $\tilde{h} : \mathbb{R}^{d_a} \rightarrow \mathbb{R}^{\tilde{d}}$ is a neural network parameterized by η_1 , and $\eta_2 \in \mathbb{R}^{\tilde{d}}$ is an additional parameter. In the language of attention, the output of $\tilde{h}(h_a(a; \theta); \eta_1)$ is described as a “key”, η_2 is a “query”, and the output of $g(a; \theta, \eta)$ is an (unnormalized) “attention weight”.

DeepRC is designed for a binary outcome or label, and employs a cross-entropy loss, corresponding to a logistic model. To define the model, we represent the sequence a_{ij} as a one-hot encoding, such that $a_{ijkl} = 1$ if the k th letter of the sequence is l and $a_{ijkl} = 0$ otherwise. Then the conditional distribution of y_i given a set of sequences $\{a_{i1}, \dots, a_{im_i}\}$ is given by,

$$Y_i \sim \text{Bernoulli} \left(\sigma \left(\gamma_a^\top \frac{\sum_{j=1}^{\tilde{m}_i} h_a(\log(c_{ij}) \tilde{a}_{ij}; \theta) g(\log(c_{ij}) \tilde{a}_{ij}; \theta, \eta) \mathbb{I}[g(\log(c_{ij}) \tilde{a}_{ij}; \theta, \eta) > Q_{0.9}(\hat{g}_i)]}{\sum_{j=1}^{\tilde{m}_i} g(\log(c_{ij}) \tilde{a}_{ij}; \theta, \eta) \mathbb{I}[g(\log(c_{ij}) \tilde{a}_{ij}; \theta, \eta) > Q_{0.9}(\hat{g}_i)]} + \gamma_y \right) \right) \quad (22)$$

where $\sigma = 1/(1 + \exp(-x))$ is the logistic function, \hat{g}_i is the empirical distribution of the attention weights

$$\hat{g}_i = \frac{1}{\tilde{m}_i} \sum_{j=1}^{\tilde{m}_i} \delta_{g(\log(c_{ij}) \tilde{a}_{ij}; \theta, \eta)},$$

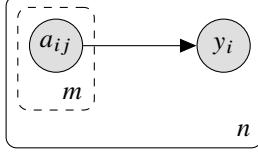


Figure S3: Causal model without confounding.

and $\mathcal{Q}_{0.9}(\hat{g}_i)$ is the value of the 90th percentile of the distribution. In the language of attention, $h_a(\log(c_{ij})\tilde{a}_{ij}; \theta)$ is the “value” and $g(\log(c_{ij})\tilde{a}_{ij}; \theta, \eta)/\sum_{j=1}^{\tilde{m}_i} g(\log(c_{ij})\tilde{a}_{ij}; \theta, \eta)$ is the “attention weight”.

The indicator $\mathbb{I}[g(\log(c_{ij})\tilde{a}_{ij}; \theta, \eta) > \mathcal{Q}_{0.9}(\hat{g}_i)]$ sets smaller attention weights to zero (note it is specific to DeepRC rather than attention-based models more broadly). The motivation for including this term is to reduce computational cost: terms with $g(\log(c_{ij})\tilde{a}_{ij}; \theta, \eta) \leq \mathcal{Q}_{0.9}(\hat{g}_i)$ can be ignored when computing the gradient of the log likelihood, reducing both time and memory requirements.

Training of DeepRC proceeds by drawing a minibatch of unique sequences $\{\tilde{a}_{i1}, \dots, \tilde{a}_{i\tilde{m}_i}\}$ uniformly for each patient repertoire, batch-normalizing the encoded sequences [Ioffe and Szegedy, 2015], and using this sample to approximate the sum over unique sequences in the numerator and denominator of Eq. 22.

DeepRC[★] We modified DeepRC slightly such that it provides causal effect estimates under a causal model with no confounding. In particular, consider the hierarchical causal model in Figure S3. After collapsing and applying σ -calculus as in Appendix A.3, we can identify the causal effect as

$$p(y; \text{do}(q_\star^a \sim \sigma_{a_\star, \epsilon})) = \int p(y | (1 - \epsilon)q^a + \epsilon\delta_{a_\star})p(q^a)dq^a \quad (23)$$

To estimate the interventional effect under this model, we need to regress the outcome on the repertoire distribution, i.e. we need to estimate a model of $p(y | q^a)$.

Although DeepRC regresses from repertoires to outcomes, it does not directly provide such a model. The key issue is that DeepRC is not a function just of the empirical distribution of sequences \hat{q}_i^a . Instead, it depends also on the total number of sequences m_i . In general, this number will depend on the experimental measurement procedure, rather than just the patient’s repertoire (recall that the true number of T cells in the repertoire is many orders of magnitude larger than the number of samples that are actually recorded by sequencing). We modify DeepRC to remove the dependence on m_i .

The DeepRC[★] model (for a binary outcome) is,

$$Y_i \sim \text{Bernoulli} \left(\sigma \left(\gamma_a^\top \frac{\mathbb{E}_{q_i^a} [h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{q_i^a} [g(A; \theta, \eta)]} + \gamma_y \right) \right) \quad (24)$$

Approximating the expectations with the empirical distribution of repertoire sequences yields,

$$\frac{\mathbb{E}_{q_i^a} [h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{q_i^a} [g(A; \theta, \eta)]} \approx \frac{\mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{\hat{q}_i^a} [g(A; \theta, \eta)]} = \frac{\sum_{j=1}^{\tilde{m}_i} c_{ij} h_a(\tilde{a}_{ij}; \theta)g(\tilde{a}_{ij}; \theta, \eta)}{\sum_{j=1}^{\tilde{m}_i} c_{ij} g(\tilde{a}_{ij}; \theta, \eta)} \quad (25)$$

Written in this way, we can see that there are two key differences from DeepRC (Eq. 22). First, instead of multiplying the sequence representation by its log counts, i.e. $\log(c_{ij})\tilde{a}_{ij}$, we multiply

the sequence’s attention weight by its counts, i.e. $g(\tilde{a}_{ij}; \theta, \eta)c_{ij}$. Second, we drop the quantile indicator function; in practice, we found the computational speedups provided by the quantile indicator to be unnecessary for training the model; this may be because we use improved GPU hardware as compared to that used in Widrich et al. [2020].

To train the model we draw a minibatch of samples from the full repertoire $\{a_{i1}, \dots, a_{im_i}\}$, rather than just the set of unique sequences, and approximate the expectations as,

$$\frac{\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{\hat{q}_i^a}[g(A; \theta, \eta)]} \approx \frac{\sum_{j \in \mathcal{S}} h_a(a_{ij}; \theta)g(a_{ij}; \theta, \eta)}{\sum_{j \in \mathcal{S}} g(a_{ij}; \theta, \eta)} \quad (26)$$

where $\mathcal{S} \subseteq \{1, \dots, m_i\}$ is the set of indices of the minibatch. We do not use batch normalization, to avoid the complexities it introduces at test time, when the model is used to estimate the effects of held-out sequences.

D Details on Semisynthetic Data

In this section we describe our procedure for constructing semisynthetic data in detail.

Let $\kappa \in \mathcal{X}$ denote a motif – a string of amino acids – of length $|\kappa|$. We define the *motif injection* function $\mathcal{I}(q; \kappa, L)$ which adds the motif κ into a sequence distribution at position L : if $\tilde{q} = \mathcal{I}(q; \kappa, L)$, then samples from \tilde{q} are generated as

$$\begin{aligned} X &\sim q \\ X_{L:L+|\kappa|} &= \kappa. \end{aligned} \quad (27)$$

That is, for each sample from q , we overwrite whatever letters were at positions $L, \dots, L + |\kappa| - 1$ with the motif κ [Widrich et al., 2020]. We also define a *motif recognition* function $D(x; \kappa)$. This function takes value 1 if the motif κ appears in the sequence, and zero otherwise, i.e.

$$D(X; \kappa) = \mathbb{I} \left[\sum_{j=1}^{|X|-|\kappa|+1} \mathbb{I}(X_{j:j+|\kappa|-1} = \kappa) > 0 \right]. \quad (28)$$

To construct the synthetic pre-selection repertoire, we start with values \tilde{q}_i^z learned from real repertoires via IGoR [Pavlović et al., 2021]. Then we inject the confounded motif κ^{con} into all patients, and the causal motif κ^{cau} into some patients. The fraction of TCRs that contain each motif is η , which we will choose to be small ($\eta \leq 0.01$).

$$\begin{aligned} \zeta_i &\sim \text{Bernoulli}(0.4) \\ q_i^z &= (1 - \eta - \eta\zeta_i)\tilde{q}_i^z + \eta\mathcal{I}(\tilde{q}_i^z; \kappa^{\text{con}}) + \eta\zeta_i\mathcal{I}(\tilde{q}_i^z; \kappa^{\text{cau}}) \end{aligned} \quad (29)$$

We take samples $z_{ij} \sim q_i^z$ to form a simulated pre-selection repertoire.

Next we generate the confounder and relative fitness. We generate

$$\begin{aligned} U_i &\sim \text{Bernoulli}(0.4) \\ r_i(x) &= \exp((6u_i - 5)D(x; \kappa^{\text{con}})) \end{aligned} \quad (30)$$

Sequences containing the confounded motif are subject to positive selection ($r_i(x) > 1$) when $u_i = 1$, and negative selection ($r_i(x) < 1$) when $u_i = 0$. The particular values of the confounder probability (0.4) and selection coefficient ($\exp(1)$ for positive selection and $\exp(-5)$ for negative

selection) are chosen such that the confounded motif will have a similar occurrence pattern in the mature repertoire as the causal motif: 40% of patients will have TCRs with the motif at prevalence $\approx \eta$, while in the rest of the patients the motif will be nearly absent.

Next we construct the mature repertoire. To do so, we use a finite sample approximation to the selection equation (Eq. 1). We first draw m' fresh samples $z'_{i1}, \dots, z'_{im'}$ from q_i^z (independent of $\{z_{ij}\}_{j=1}^{m_i}$) and then take

$$q_i^a = \frac{1}{\sum_{j=1}^{m'} r_i(z'_{ij})} \sum_{j=1}^{m'} r_i(z'_{ij}) \delta_{z'_{ij}} \quad (31)$$

We sample the mature repertoire sequences as $a_{ij} \sim q_i^a$ for $j \in \{1, \dots, m_i\}$.

Finally, the outcome variable y depends on (a) the presence of a subpopulation of TCRs with the causal motif, and (b) the confounder:

$$Y_i \sim \text{Normal}(\gamma_a \mathbb{I}(\mathbb{E}_{q_i^a}[D(x, \kappa^{\text{cau}})] > \eta/2) + \gamma_u u_i + \gamma_0, \tau_y). \quad (32)$$

We focus on the regime where γ_u is substantially larger than γ_a ($\gamma_u = 2$ versus $\gamma_a = 0.4$). This creates the impression, if confounding is ignored, that sequences with the confounded motif κ^{con} have a larger effect than sequences with the true causal motif κ^{cau} . We set the standard deviation τ_y to be small compared to γ_u and γ_a ($\tau_y = 0.1$). Previous simulation studies of repertoire classification have focused on a regime with no noise, so using low values of τ_y ensures our results are at least somewhat comparable to these previous studies.

We used the real repertoire data from [Emerson et al. \[2017\]](#) (<https://clients.adaptivebiotech.com/pub/emerson-2017-natgen>) as the basis for our semisynthetic data. The dataset contains 786 patients. There is an average of 31,000 sequences per patient (minimum 177, maximum 183,000). The pre-selection repertoire \tilde{q}_i^z is estimated from the non-productive sequences via IGoR, following the protocol in Appendix G.4. We use motifs of length $|\kappa^{\text{cau}}| = |\kappa^{\text{con}}| = 3$, and inject them at position $L = 3$. To ensure that the motifs are not very common or extremely rare in existing sequences, we first sort existing length 3 subsequences in the initial repertoires by frequency. We then choose κ^{cau} and κ^{con} at random from among those subsequences whose frequency is between the 10th and 20th percentile of all subsequences. We work only with the CDR3 region of the TCR β , starting immediately after the cysteine on 5' side and ending immediately before the phenylalanine on the 3' side.

It is important to note that in our semisynthetic experiments, we train CAIRE's fitness model directly on simulated productive sequences from the pre-selection repertoire $\{z_{ij}\}_{j=1}^{m_i}$, rather than run inference with IGoR on nonproductive sequences and then draw productive sequences from the estimated model. This is because the underlying parametric model used by IGoR cannot describe the motif-injected repertoire \tilde{q} ; we are also limited by the fact that inference in IGoR is computationally intensive. Hence, our semisynthetic experiments should be understood as an evaluation of the CAIRE model alone (Eqs. 19 and 20), and not in combination with IGoR.

E Details on Architectures and Training

In this section we detail CAIRE's model architecture and training, as well as that of other methods we compare to in experiments.

E.1 Models

We consider the following models.

CAIRE The CAIRE model (Eqs. 19 and 20), restated, is

$$s_{ij} \sim \text{Bernoulli}(\sigma(\rho_i^\top h_r(x_{ij}; \phi) + \beta_i)) \quad (33)$$

$$y_i \sim \text{Normal}\left(\gamma_a^\top (\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] - W \cdot \rho_i - B) + \gamma_r^\top \rho_i + \gamma_y, \tau_y\right) \quad (34)$$

along with the separate propensity model,

$$\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] \sim \text{Normal}(W \cdot \rho_i + B, \tau_e). \quad (35)$$

No propensity CAIRE This model is a variant of CAIRE without the propensity correction.

$$s_{ij} \sim \text{Bernoulli}(\sigma(\rho_i^\top h_r(x_{ij}; \phi) + \beta_i)) \quad (36)$$

$$y_i \sim \text{Normal}\left(\gamma_a^\top \mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] + \gamma_r^\top \rho_i + \gamma_y, \tau_y\right). \quad (37)$$

Uncorrected This model does not correct for confounding.

$$y_i \sim \text{Normal}\left(\gamma_a^\top \mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)] + \gamma_y, \tau_y\right). \quad (38)$$

DeepRC[★] This model is a variant of DeepRC [Widrich et al., 2020], adapted for causal effect estimation (Appendix C). Since the outcome variable is continuous, we use a Gaussian outcome distribution.

$$y_i \sim \text{Normal}\left(\gamma_a^\top \frac{\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{\hat{q}_i^a}[g(A; \theta, \eta)]} + \gamma_y, \tau_y\right) \quad (39)$$

Attention CAIRE This model combines CAIRE with the attention-based model parameterization from DeepRC[★].

$$s_{ij} \sim \text{Bernoulli}(\sigma(\rho_i^\top h_r(x_{ij}; \phi) + \beta_i)) \quad (40)$$

$$y_i \sim \text{Normal}\left(\gamma_a^\top \left(\frac{\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{\hat{q}_i^a}[g(A; \theta, \eta)]} - W \cdot \rho_i - B\right) + \gamma_r^\top \rho_i + \gamma_y, \tau_y\right), \quad (41)$$

along with the separate propensity model,

$$\frac{\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{\hat{q}_i^a}[g(A; \theta, \eta)]} \sim \text{Normal}(W \cdot \rho_i + B, \tau_e). \quad (42)$$

E.2 Architecture details

In this section we detail the architecture of the neural networks in the models.

- $h_a(\cdot; \theta)$: This network extracts features from TCR CDR3 sequences, for estimating causal effects. Its architecture follows the feature extractors in DeepRC [Widrich et al., 2020].

We first encode each sequence as follows: (1) we append to the amino acid sequence an end token, marking the final position, (2) we one-hot encode the entire sequence, such that $a_{ijkl} = 1$ if the amino acid in the k th position is l and $a_{ijkl} = 0$ otherwise, and (3) we pad the one-hot encoded sequence with zeros out to the maximum sequence length observed in the

dataset, i.e. $a_{ijkl} = 0$ for all k greater than the sequence length, and (4) we add three position features describing the sequence start, end, and center, as described in detail in [Widrich et al. \[2020\]](#), Appendix A2. The fully encoded sequence a_{ij} is a matrix of size $L_{\max} \times 24$, where L_{\max} is the maximum length of sequences in the dataset, and $24 = (20 \text{ amino acids}) + (1 \text{ end token}) + (3 \text{ position features})$.

Given a sequence in this encoding, the network h_a applies a 1D convolutional neural network (CNN) with a SELU nonlinearity [[Klambauer et al., 2017](#)], then takes the maximum across positions: $h_a(a_{ij}; \theta)_\ell = \max_k(\text{SELU}(\text{CONV1D}(a_{ij}; \theta)_{k\ell}))$, where CONV1D is the 1D convolution function. The resulting output is in \mathbb{R}^{d_a} , with d_a corresponding to the number of channels in the CNN. The trainable parameters θ are the weights and biases of the CNN, i.e. the convolutional filters. The convolution kernel size ζ (measured in number of amino acids) is a hyperparameter.

- $h_r(\cdot; \phi)$: This network extracts features from TCR CDR3 sequences, for estimating fitness. Its first layer has an identical architecture to h_a , that is it can be written in the form $h_r(a_{ij}; \phi) = \tilde{h}_r(h_a(a_{ij}; \phi_0); \phi_1)$. (The parameters are not shared with h_a , however, i.e. $\phi_0 \neq \theta_0$.) The remaining layers of h_r are feedforward: \tilde{h}_r is a feedforward neural network with SELU nonlinearities.
- $g(a; \theta, \eta)$: This network is used in the models DeepRC * and Attention CAIRE, and outputs a scalar weight in \mathbb{R}_+ . Its architecture follows the attention network used in DeepRC [[Widrich et al., 2020](#)]. In particular, the first layer is identical to, and shares parameters with, h_a . More precisely, we can write $g(a_{ij}; \theta, \eta) = \exp(\tilde{g}(h_a(a; \theta); \eta))$. The remaining layers are feedforward: $\tilde{g}(\cdot; \eta)$ is a two-layer feedforward neural network with SELU nonlinearities.

E.3 Amortized inference

The selection representation ρ_i and the offset term β_i are per-patient latent variables. To achieve scalable, stochastic gradient-based inference on large datasets, we amortize inference of ρ_i and β_i across patients. The inference network is parameterized as,

$$(\rho_i, \beta_i) = \text{enc}(\{(x_{ij}, s_{ij})\}_{j=1}^{2m_i}; \theta', \lambda) = \tilde{e}(\mathbb{E}_{\hat{q}_i^a}[h_a(A; \theta')] - \mathbb{E}_{\hat{q}_i^z}[h_a(A; \theta')]; \lambda), \quad (43)$$

where $\tilde{e}(\cdot; \lambda)$ is a neural network. The neural network h_a has the architecture described in Appendix E.2 (but $\theta' \neq \theta$). This encoder parameterization is intended to help the model focus on differences between the pre-selection and post-selection repertoires, by taking the difference between the embedding of q_i^a (estimated from mature repertoire) and the embedding of q_i^z (estimated from the pre-selection repertoire). The function \tilde{e} is a single-layer feedforward neural network with SELU nonlinearity, i.e. it takes the form $\tilde{e}(\cdot; \lambda) = \text{Linear}_{\lambda_1}(\text{SELU}(\text{Linear}_{\lambda_0}(\cdot)))$ where Linear denotes a linear layer. The parameters $\lambda = (\lambda_0, \lambda_1)$ determine the weights and biases of the linear layers.

In preliminary experiments, we compared to an alternative amortization network that pools the pre-selection and mature repertoire sequences, rather than take the difference in their representation. We found that the parameterization in Eq. 43 encourages the model to focus on features that distinguish between the pre-selection and post-selection repertoire, and helps avoid poor solutions where ρ_i contains additional information about q_i^a beyond the relative fitness r_i .

Our inference network provides an estimate of the maximum likelihood value of ρ_i and β_i . We also explored amortized variational inference of ρ_i and β_i , following the logic of variational

autoencoders [Kingma and Welling, 2019]. However, we found convergence to be quite poor in preliminary experiments. Moreover, since there are many sequences per patient, uncertainty in the per-patient latent variables ρ_i and β_i appeared to be low. We therefore did not pursue this variational inference approach further.

E.4 Effect estimates

After training, we can compute the estimated effect using Eq. 4 and the empirical distribution of $p(q^a, r)$, or more precisely, the empirical distribution of the estimates \hat{q}_i^a and ρ_i . For CAIRE, the estimated effect is,

$$\text{ATE}(a_\star, \epsilon) \approx \frac{1}{n} \sum_{i=1}^n \gamma_a^\top (\mathbb{E}_{(1-\epsilon)\hat{q}_i^a + \epsilon\delta_{a_\star}} [h_a(A; \theta)] - W \cdot \rho_i - B) + \gamma_r^\top \rho_i + \gamma_y \quad (44)$$

$$- \frac{1}{n} \sum_{i=1}^n \gamma_a^\top (\mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)] - W \cdot \rho_i - B) + \gamma_r^\top \rho_i + \gamma_y \quad (45)$$

$$= \epsilon \gamma_a^\top \left(h_a(a_\star; \theta) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)] \right) \quad (46)$$

For the No propensity CAIRE model and the Uncorrected model, the estimated effect works out to the same expression.

For DeepRC * and the Attention CAIRE model, the estimated effect becomes,

$$\begin{aligned} \text{ATE}(a_\star, \epsilon) \approx & \frac{1}{n} \sum_{i=1}^n \gamma_a^\top \left(\frac{(1-\epsilon)\mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)g(A; \theta, \eta)] + \epsilon h_a(a_\star; \theta)g(a_\star)}{(1-\epsilon)\mathbb{E}_{\hat{q}_i^a} [g(A; \theta, \eta)] + \epsilon g(a_\star; \theta, \eta)} \right. \\ & \left. - \frac{\mathbb{E}_{\hat{q}_i^a} [h_a(A; \theta)g(A; \theta, \eta)]}{\mathbb{E}_{\hat{q}_i^a} [g(A; \theta, \eta)]} \right). \end{aligned} \quad (47)$$

E.5 Training details

We use early stopping for regularization, with a validation set consisting of 12.5% of the data. As our stopping condition we use a score measuring the predictive performance of CAIRE. A typical choice of early stopping score is the training objective; however, since there are large number of sequences in each individual’s repertoire, the overall value of the training objective is dominated by the contribution of the selection model (Eq. 19, line 1), and the contribution of the outcome model (Eq. 19, line 2) is washed out. To create a more balanced score, we heuristically combine performance metrics for the selection model and the outcome model, which are each between zero and one. In particular, we sum (1) the accuracy of the selection model in predicting s and (2) the coefficient of determination (R^2) of the outcome model with y . We compute this score on the validation set after every 500 iterations, and take the model corresponding to the best recorded value over the entire training run.

We place priors on the main parameters of the model, γ , ρ_i , β_i , W , B and τ (Appendix F.2). Following previous work on autoencoders, the prior on ρ_i and β is annealed during training: the log likelihood of the prior is weighted by a factor ξ_t , which increases linearly from 0 to 1 [Fu et al., 2019].

The sequence feature extraction network $h_a(\cdot; \theta)$ is computed in low precision, using bfloat16 [Kalamkar et al., 2019]. This occurs wherever the h_a function is used, i.e. not just in the model of y_i , but also in the first layer of h_r and in g . All other computations are at single precision, float32.

All models were implemented in the PyTorch-based probabilistic programming language Pyro (version 1.8.4) [Bingham et al., 2019]. All models were trained using an NVIDIA A100 GPU with 80GB memory. All models were trained for the same, fixed amount of wall-clock time, 5 minutes. This was chosen based on preliminary experiments suggesting that it was sufficient for convergence (according to the diagnostic of Pesme et al. [2020]) across the different model classes.

F Details on Semisynthetic Experiments

F.1 Evaluation details

For evaluation, we hold out data from 12.5% of patients. We then look at the repertoires $\{a_{ij}\}_{j=1}^m$ of each held-out patient who had the causal motif injected, i.e. each patient i with $\zeta_i = 1$ (Appendix D). We examine the effect of interventions that use each of their repertoire sequences, i.e. $\widehat{\text{ATE}}(a_\star = a_{ij}, \epsilon = 0.01)$ for $j \in \{1, \dots, m\}$. We then evaluate how well this effect estimate can discriminate causal sequences, $\{a_{ij} : D(a_{ij}; \kappa^{\text{cau}}) = 1\}$, from non-causal sequences, $\{a_{ij} : D(a_{ij}; \kappa^{\text{cau}}) = 0\}$. Performance is quantified with the area under the precision-recall curve (PR-AUC). The final performance of the model is the average PR-AUC across the held-out motif-injected patients.

To evaluate the statistical significance of average differences between model performance, we use a permutation-based t-test (scipy `ttest_ind` with 10000 permutations).

F.2 Hyperparameters

In this section we describe our handling of hyperparameters in the semisynthetic experiments. Key hyperparameters (especially those governing the overall dimensionality of a model) were optimized for each method that we compare to, on each dataset. Due to computational cost, optimizing all hyperparameters was infeasible; the remaining hyperparameters were fixed based on preliminary experiments, and based on previous work on DeepRC [Widrich et al., 2020] and SIN [Kaddour et al., 2021].

Hyperparameter optimization was performed using BoTorch [Balandat et al., 2020], via the Ax interface (<https://github.com/facebook/Ax>, version 0.3.4). For CAIRE’s search space, Ax’s default strategy consisted of a Sobol sequence for six iterations, followed by Bayesian optimization with a Gaussian process and the expected improvement acquisition function. We use 10 rounds of hyperparameter optimization for all experiments.

F.3 Experiments

Model comparison In this experiment, we compare the methods in Appendix E.1. We draw 15 independent semisynthetic datasets, generated as in Appendix D (with motif injection rate $\eta = 0.01$). We then evaluate each method on each dataset.

Low motif rates In this experiment we compare CAIRE to the Uncorrected method (Appendix E.1). We evaluate on semisynthetic datasets in which the motif injection rate η is set to 0.001, 0.005, and 0.01. For each value of η , we sample 10 independent semisynthetic datasets, and evaluate each method on each dataset.

No confounding In this experiment we compare CAIRE to the Uncorrected method (Appendix E.1). We sample 10 independent semisynthetic datasets in which confounding does not

Table 2: Hyperparameters for semisynthetic experiments. Hyperparameters with a search range are optimized over that range using Bayesian optimization. Remaining hyperparameters are fixed at the given value.

Hyperparameter	Value or search range
Dimension d_r of the selection representation ρ_i	[2, 32]
Dimension d_a of the repertoire representation $\mathbb{E}_{q_i^a} [h_a(A; \theta)]$	[8, 32]
Kernel size of CNNs in h_a , h_r , and enc	{5, 7, 9}
Number of channels of the CNNs in h_r and enc	8
Number of layers in the selection network \tilde{h}_r	3
Dimension of hidden layers in the selection network \tilde{h}_r	16
Number of layers in the attention network \tilde{g}	2
Dimension of hidden layers in the attention network \tilde{g}	32
Prior on entries of γ_a , γ_r and γ_y	Normal(0, 100)
Prior on entries of ρ_i	Normal(0, 1)
Prior on β_i	Normal(0, 10)
Prior on entries of W and B	Normal(0, 10)
Prior on τ_y and τ_e	LogNormal(-1, 2)
Dimension of hidden layer in the encoder network \tilde{e}	8
Batch size: patients	8
Batch size: sequences per patient	16,384
Learning rate	0.01
Weight decay	0.01
Validation set size	12.5% of patients
Test set size	12.5% of patients
Training time	5 minutes
Annealing time	3 minutes

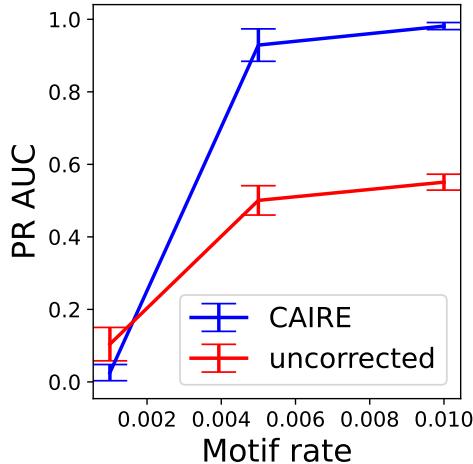


Figure S4: Performance on semisynthetic data with increasing motif injection rate η .

contribute to the outcome, i.e. with $\gamma_u = 0$. We set the motif injection rate η to 0.01. We then evaluate each model on each dataset.

G Details on COVID Severity Data

Snyder et al. [2020] and Nolan et al. [2020] constructed a TCR sequencing dataset from a large cohort of individuals currently or previously infected with SARS-CoV-2. They also record clinical information about the patients, along with demographic information. Details on the data, including the sequencing procedure and the patient populations, can be found in the original publications. Here, we describe how we preprocess this data to study the causal impacts of TCRs on disease. The data is available at <https://clients.adaptivebiotech.com/pub/covid-2020>. The dataset we used was dated to July 2020, and we accessed it January 2024.

G.1 Patient Outcomes

We first constructed an outcome score for each patient measuring their overall disease severity. The raw data was collected from several different patient cohorts at different study sites, and contains a heterogenous mixture of clinical information, with different variables missing for different patients. We sought to construct a single summary of severity. First, we determined whether a patient was hospitalized. Patients with `hospitalized == True`, `days_in_hospital > 0` or `covid_unit_admit == True` were labeled as *hospitalized*. Those patients with `hospitalized`, `days_in_hospital` and `covid_unit_admit` all missing were labeled as *missing hospitalization*. All other patients were labeled as having *mild disease*. Next, we determined whether a patient had severe disease. Patients with `icu_admit == True` or `death == True` were labeled as having severe disease (regardless of the previous label). All remaining *hospitalized* patients were labeled as having *moderate disease*. All remaining *missing hospitalization* patients were labeled as *missing outcome*. Patients with missing outcome data were dropped from further analysis.

The data contained a limited number of instances of repeated repertoire samples taken from the same patient (14% of all repertoire sequencing samples). In instances where the clinical information corresponding to each sample differed, we used the more severe patient outcome. E.g. if a patient is first hospitalized, and then dies, we label them as having *severe disease*, and use the repertoire sequencing data labeled with this clinical outcome. When the outcome information was the same across different samples from a patient, we use the repertoire sequencing data corresponding to the earliest doctor visit, i.e. the time-point closest to disease onset. When multiple repertoire samples come from the same visit of the same patient, or information on time-points is missing, we take the repertoire sample with the largest number of TCR sequences.

G.2 Patient Demographics

The final dataset consisted of $n = 507$ patients, of which 86 had severe disease (patients who were in the ICU or died), 187 had moderate disease (patients who were hospitalized) and 234 had mild disease (patients who were neither hospitalized nor died).

Among patients without missing demographic data, we found the average age was 56.9 (standard deviation 18.2), and 48% were male. 84.4% were non-Hispanic Caucasian, 9.0% Hispanic, 1.8% Asian or Pacific Islander, and 1.8% Black or African American, with remaining racial and ethnic groups below 1%.

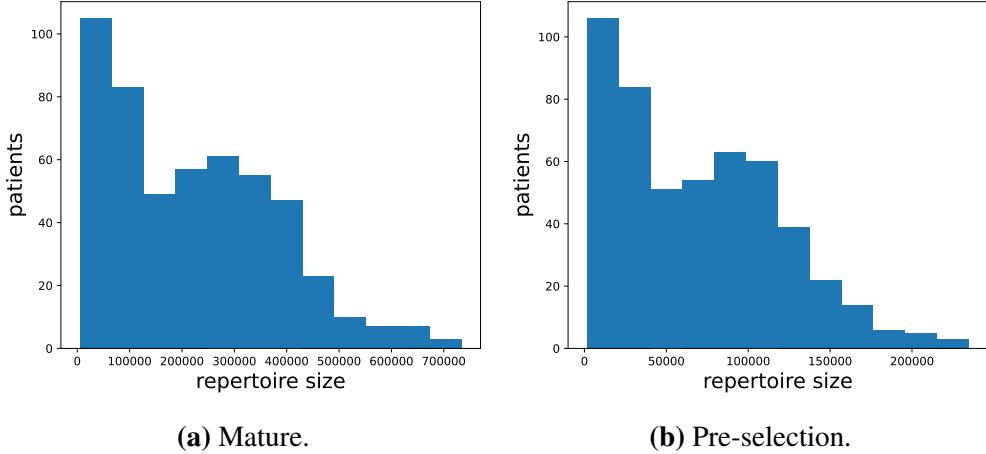


Figure S5: Distribution of repertoire sizes across patients. Note the pre-selection repertoires are simulated. The histogram bin size is set according to the Freedman-Diaconis rule [Freedman and Diaconis, 1981].

G.3 Mature Repertoires

For each patient we assemble a dataset of mature TCR sequences. The immunoSEQ assay used for sequencing in [Snyder et al. \[2020\]](#) is a bulk RNA sequencing assay, which records a segment of the TCR covering the TCR β CDR3. For the mature repertoire, we take only productive sequences (`frame_type == 'In'`), excluding those with a stop codon or frameshift mutation. We work with amino acids, taking the CDR3 region starting immediately after the conserved cysteine (C) on 5' side and ending immediately before the conserved phenylalanine (F) on the 3' side. The distribution of mature repertoire sizes (in terms of number of unique sequences) is shown in Figure S5a. The mean CDR3 length across all sequences was 12.5 amino acids (standard deviation 1.8), excluding the terminal C and F. We also record the frequency of each sequence (`productive_frequency`), which is an estimate of how often it occurs in the population of T cells.

G.4 Immature Repertoires and IGoR

We estimate each patient's immature TCR repertoire using non-productive sequence data, via IGoR [[Marcou et al., 2018](#)].

For each patient, we first extract the non-productive sequences (`frame_type != 'In'`), which have frameshift or truncation mutations. We record the 65 nucleotides at the 3' end of each sequence (in immunoSEQ, priming is done in the constant region of the TCR, on the 3' side).

We then train an IGoR model on each patient's non-productive sequences. Due to the computational limitations of IGoR, patient repertoires with more than 10,000 sequences were randomly subsampled down to 10,000 sequences, and inference over rearrangement scenarios was done with a “Viterbi-like” algorithm (using the `--MLSO` flag). We use the default IGoR model architecture. We also use the default reference set of human V(D)J genes; note that our analysis is therefore limited in that it does not account for possible germline allelic variation in these genes, beyond the variation present in IGoR’s reference set [[Slabodkin et al., 2021](#)]). IGoR reports low sequencing error rates across all patients, suggesting its inferences are reasonable (mean error across patients: 0.0052, 95th percentile: 0.0061, max: 0.0072)

For the pre-selection repertoire to be an effective instrumental variable, it must vary across patients. We therefore examined variation across patients in the pre-selection repertoire distribution

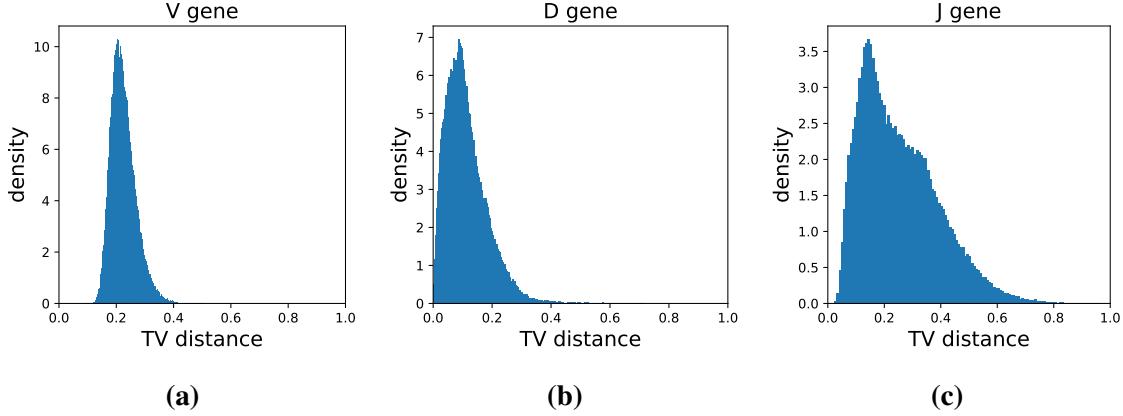


Figure S6: Distribution of gene usage differences among pairs of patients. The histogram bin size is set according to the Freedman-Diaconis rule [Freedman and Diaconis, 1981].

estimated by IGoR. In particular, we computed the total variation distance between the marginal distribution over V gene usage, among all pairs of individuals (that is, we look at the absolute value of the difference in estimated V gene frequency, summed over all genes). The distribution exhibits a peak away from zero (Figure S6a). The D and J gene distributions show similar patterns, with a wide range of distances between patients Figures S6b and S6c. So, based on these estimates, there is indeed variation among patients in the pre-selection repertoire [Slabodkin et al., 2021].

Once the IGoR model is trained, we use it to draw samples from the distribution of the pre-selection mature repertoire. In particular, we first sample a number of sequences equal to the total number of TCR sequences measured in the patient. We then retain only those sampled sequences that are productive (`is_inframe == True`, `anchors_found == True`, and no internal stop codon). In other words, we use rejection sampling to obtain samples from the pre-selection distribution over productive sequences. We translate the nucleotide sequences into amino acids, and take the CDR3 region from just after the 5' cysteine to just before the 3' phenylalanine, as with the mature repertoire sequences. The distribution of generated pre-selection repertoire sizes (in terms of number of sequences) is shown in Figure S5b. The mean CDR3 length across all sequences from all patients was 13.2 amino acids (standard deviation 2.3).

G.5 MIRA data

Snyder et al. [2020] and Nolan et al. [2020] also collect a dataset of TCRs that bind SARS-CoV-2 epitopes, using MIRA assays [Klinger et al., 2013, 2015]. In these experiments, patient T cells are screened against viral peptides, and TCR sequences from cells that bind each epitope are recorded. The epitopes were delivered exogenously, as pools of synthetic peptides. The epitope sequences were chosen from among those viral peptides predicted by NetMHCpan to bind common class I and class II MHC HLA genes [Nielsen et al., 2003, Andreatta and Nielsen, 2016]. In some experiments overlapping or nearby peptides were pooled together, in which case one cannot distinguish precisely which epitope a TCR has bound.

Our aim is to use this data to evaluate how well causal estimates from CAIRE predict *in vitro* binding. Following the strategy described in Appendix H, we will compare binding TCRs from patients to repertoire sequences drawn from the same patient. For this reason, we focus just on MIRA experiments performed on patients for which we also have repertoire sequencing data.

We ignore unproductive TCR sequences (those with internal stop codons, etc.), under the

assumption that they represent false positives. We also focus just on MIRA data from a panel of class I MHC epitopes or a panel of class II MHC epitopes, dropping the smaller-scale “minigene” experiments, which do not provide exact knowledge of the epitope.

The final dataset contains a total of 32,770 binders and 11 million unlabeled repertoire sequences, drawn from 71 patients. Binding assays with class I MHC epitopes were performed on 66 patients, and assays with class II were performed on 6 patients; 28,640 TCR sequences were found that bound class I in total, while 4,130 bound class II. The average number of binders per patient was 462, with a minimum of 12 and a maximum of 2,344. The average number of repertoire sequences per patient was 151,000, with a minimum of 13,910 and a maximum of 565,000. There were 265 class I epitopes with at least one TCR binder, and 56 class II epitopes.

H In vitro Binding Comparison Method

In this section we detail our strategy for comparing causal effect estimates to *in vitro* TCR binding data. The key challenge is that the binding assay (Appendix G.5) only provides data on the sequences of TCRs that bind a target, and does not provide data on the sequences of TCRs that do not bind the target. In other words, it provides only positive examples of binders, and not negative examples of non-binders. To evaluate our models, we will therefore combine the MIRA data with unlabeled data from unbiased repertoire sequencing. In this section, we formalize our evaluation mathematically.

We first describe the data generating process. We assume that TCR sequences are drawn from some underlying distribution $q_0(x)$, corresponding to the distribution of TCRs within a patient or group of patients. We further assume that the probability that a given TCR x binds the antigen is $q_0(d | x)$. Here, $d \in \{0, 1\}$ is a binary variable, with $d = 1$ indicating binding.

We have access to unlabeled sequences, collected via repertoire sequencing. We assume that this data is drawn from the underlying TCR distribution,

$$\tilde{x}_{1:\tilde{n}} \stackrel{i.i.d.}{\sim} q_0(x). \quad (48)$$

We also have data on sequences that bind the antigen, which comes from screening patient T cells. We assume that this data is drawn as,

$$x_{1:n} \stackrel{i.i.d.}{\sim} q_0(x | d = 1). \quad (49)$$

This says that the sequences collected in the binding assay come from (1) drawing samples from q_0 , and then (2) filtering to retain just samples that bind the target antigen ($d = 1$).

Our goal is to understand how well a function $g(x)$ predicts binding. For example, $g(x)$ can be the treatment effect estimate from CAIRE, $g(x) = \widehat{\text{ATE}}(x, \epsilon)$. We will focus on the area under the receiver operating characteristic curve (ROC AUC) as our performance metric. Let (x, d) and (x', d') be independent random variables distributed according to $q_0(x, d)$. The ROC AUC is,

$$\text{A}(g) = \mathbb{P}_{q_0}(g(x) \geq g(x') | d = 1, d' = 0). \quad (50)$$

In other words, the ROC AUC is the probability that a random positive example will be scored higher than a random negative example. Standard methods for estimating the ROC AUC rely on samples from $q_0(x | d = 0)$, that is, negative examples. The challenge here is that negative examples are unavailable.

Instead, we will evaluate how well $g(x)$ discriminates positive examples from unlabeled examples, and use this evaluation to produce an estimate of the ROC AUC. In particular, we consider the ROC AUC for discriminating positive labeled data from unlabeled data,

$$\tilde{A}(g) = \mathbb{P}_{q_0}(g(x) \geq g(x') \mid d = 1). \quad (51)$$

This quantity can be estimated by applying a standard ROC AUC estimator to positive labeled data (binders) and unlabeled data (repertoire sequencing data). It is related to $A(g)$ as,

$$\begin{aligned} \tilde{A}(g) &= \mathbb{E}_X[\mathbb{E}_{X'}[\mathbb{I}(g(X) \geq g(X'))] \mid d = 1] \\ &= \mathbb{E}_X[\mathbb{E}_{X'}[\mathbb{I}(g(X) \geq g(X')) \mid d' = 1]q_0(d = 1) + \mathbb{E}_{X'}[\mathbb{I}(g(X) \geq g(X')) \mid d' = 0]q_0(d = 0) \mid d = 1] \\ &= 0.5 q_0(d = 1) + A(g)q_0(d = 0). \end{aligned} \quad (52)$$

where the last line comes from linearity of expectation and from the fact that the ROC AUC for discriminating two variables drawn from the same distribution is 0.5. We can see that the ROC AUC for discriminating unlabeled examples depends on the ROC AUC for discriminating negative examples.

In practice, we do not have direct data on $q_0(d = 1)$. However, [Snyder et al. \[2020\]](#) estimate, based on various lines of evidence, that roughly 0.2% of the repertoire is involved in binding SARS-CoV-2, so roughly speaking we expect $q_0(d = 1) \approx 0.002$ (Figure 4). This would make $\tilde{A}(g)$ a close approximation of $A(g)$. Even if this estimate of $q_0(d = 1)$ is inexact, however, note that \tilde{A} still provides a conservative underestimate of $A(g)$. Moreover, if a predictor g' has better performance than g at discriminating binders, i.e. $A(g') > A(g)$, then the predictor will have better performance at discriminating unlabeled data, i.e. $\tilde{A}(g') > \tilde{A}(g)$, regardless of the value of $q_0(d = 1)$. We can therefore use \tilde{A} to compare binding predictors.

Note it is important in the derivation in Eq. 52 that the unlabeled data and the labeled data both come from the same distribution q_0 . If the unlabeled data came from a different distribution – say, repertoire sequences drawn from a different group of patients – then Eq. 52 need not hold.

In summary, we focus on $\tilde{A}(g)$ as an evaluation metric for how well effect estimates predict binding. We report the empirical ROC AUC for discriminating binders (measured by MIRA) from unlabeled sequences (measured by repertoire sequencing).

I Details on COVID Severity Model

I.1 Hyperparameters and training

The repertoire sequencing data consists of a set of unique sequences along with their weights, which correspond to an estimate of the sequence's frequency in the population of T cells under study. This gives the empirical distribution estimate of the mature repertoire distribution of the form: $\hat{q}_i(a) = \sum_{j=1}^{m_i} w_{ij}\delta_{a_{ij}}(a)$, where the repertoire sequences are $\{a_{i1}, \dots, a_{im}\}$ and their corresponding weights are $\{w_{i1}, \dots, w_{im}\}$ (these weights are normalized to one). During training, we sample minibatches of sequences by drawing samples from $\hat{q}_i(a) = \sum_{j=1}^{m_i} w_{ij}\delta_{a_{ij}}(a)$.

We use the same hyperparameter settings as in the semisynthetic experiments (Appendix F.2), except we increase the training time to 10 minutes. After preliminary optimization, we fix the dimension of the selection representation and the repertoire representation at $d_r = d_a = 32$, and the kernel size at 9. That is, we do not perform Bayesian optimization. For the Uncorrected method and Attention CAIRE method, we use the same hyperparameter settings.

I.2 Non-neural CAIRE

As an additional comparison method, we evaluated a variant of CAIRE in which the non-linear, neural-network-based components were stripped away, and replaced with low-dimensional linear components and biophysical inductive biases. In detail, we modify the model specifications in Appendix E.2 as follows:

- $h_a(\cdot; \theta)$: Instead of one-hot encoding, amino acids are encoded with their corresponding row in the BLOSUM50 substitution matrix, so that amino acids with similar biophysical properties receive similar encodings [Henikoff and Henikoff, 1992]. The nonlinearity and max-pooling are removed, leaving a linear convolution: $h_a(a_{ij}; \theta)_\ell = \sum_k \text{CONV1D}(a_{ij}; \theta)_{k\ell}$.
- $h_r(\cdot; \phi)$: Instead of a feedforward neural network, we use a single linear layer.

The hyperparameters follow (Appendix F.2), except we fix the dimension of the selection representation to a small value ($d_r = d_a = 4$) and fix the size of the kernel to 3. Training time is set to 10 minutes.

I.3 Held-out sequences

In Section 5.1, as held-out sequences, we use the 11 million repertoire sequences collected from the patients the MIRA experiments were performed on. (These are not the sequences found in the binding assay, but rather those collected via unbiased repertoire sequencing.)

I.4 Effect uncertainty

To construct estimates of uncertainty, we interpret the ensemble of 24 models as approximate samples from the Bayesian posterior of the CAIRE model [Lakshminarayanan et al., 2017, Wilson and Izmailov, 2020]. For each candidate sequence a_\star and dosage ϵ , we compute the effect estimated provided by each model in the ensemble, $\widehat{\text{ATE}}_1, \dots, \widehat{\text{ATE}}_K$ (where $K = 24$ since there are 24 models in the ensemble). We fit a Gaussian to this data, obtaining $\text{Normal}(\hat{\mu}, \hat{\sigma})$ where $\hat{\mu}$ and $\hat{\sigma}$ are the empirical mean and standard deviation respectively. Here, the average $\hat{\mu} = \frac{1}{K} \sum_{k=1}^K \widehat{\text{ATE}}_k$ provides a point estimate for the effect, i.e. it is an estimate of the posterior mean. Moreover, since each $\widehat{\text{ATE}}_k$ is an approximate sample from the posterior over the effect, the distribution $\text{Normal}(\hat{\mu}, \hat{\sigma})$ provides a rough approximation to the full posterior, $p(\text{ATE}(a_\star, \epsilon) | \mathcal{D})$, where \mathcal{D} denotes the dataset. As a measure of significance, we estimate the posterior probability that the sign of the effect is the opposite of the sign of $\hat{\mu}$, namely $\tilde{p} = p_{\text{Normal}}(x \leq 0 | |\hat{\mu}|, \hat{\sigma})$.

I.5 Dosage calculation

We use $\epsilon = 0.1$ in all the effect estimates we report. We chose this value based on rough calculation of the dosage of existing TCR-based therapies. In a review of clinical trials of TCR-engineered T cell therapy, Baulu et al. [2023] report several trials that use dosages as high as 10^{11} cells per patient. There are roughly 10^{12} T cells in an adult human [Lythe et al., 2016]. This suggests that it is tractable to intervene on TCR repertoires such that about 10% of the repertoire is a chosen sequence, giving $\epsilon = 0.1$.

Table 3: Predictive performance of COVID models. The first column gives the coefficient of determination between the model’s predictions and the outcome, while the second and third columns give the variance explained by the treatment and confounder terms respectively. All values are calculated on heldout data. For each, we report the median value across the model ensemble, along with a 95% confidence interval of the median (calculated based on the quantiles of the binomial distribution). We use the median as a robust statistic, due to outliers in the data; see Figure S7 for the raw data.

	Outcome R^2	Treatment var. explained	Confounder var. explained
CAIRE	0.035 [0.022, 0.059]	0.012 [0.00, 0.05]	0.026 [0.01, 0.06]
Attention CAIRE	0.032 [0.010, 0.062]	0.032 [0.01, 0.04]	0.024 [0.00, 0.04]
Non-Neural CAIRE	0.024 [0.011, 0.036]	0.003 [0.00, 0.01]	0.022 [0.00, 0.04]
Uncorrected	0.031 [0.014, 0.046]	0.031 [0.02, 0.05]	0.000 [0.00, 0.00]

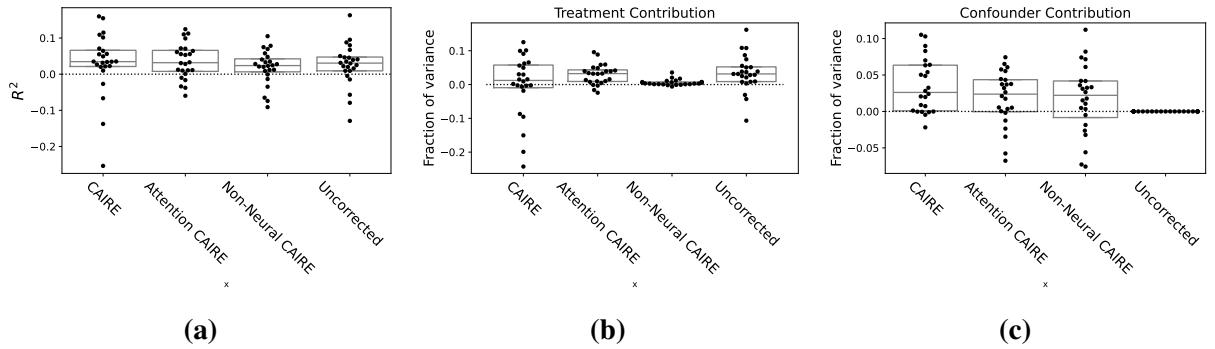


Figure S7: Predictive performance across the COVID model ensemble. Distribution across the model ensemble of the coefficient of determination R^2 with the outcome (Figure S7a), variance in the outcome explained by the treatment term (Figure S7b) and variance in the outcome explained by the confounder term (Figure S7c).

I.6 Outcome predictive performance

Table 3 and Figure S7 report the performance of CAIRE and comparison models at predicting the outcome y on held-out data. We find CAIRE, its version with attention, and the uncorrected method show similar performance. Non-Neural CAIRE performs considerably worse. Table 3 and Figure S7 also report the variance in y explained by the treatment term of CAIRE, namely $\gamma_a^\top (\mathbb{E}_{\hat{\theta}_i^q} [h_a(A; \theta)] - W \cdot \rho_i - B)$, and the variance explained by the confounder term, $\gamma_r^\top \rho_i$. Both are also calculated on held-out data.

I.7 Confounder representation

We sought to determine whether the latent representation of fitness learned by CAIRE, ρ_i , contained information about other patient covariates. In particular, we investigated whether the learned representation contained information about patient age, gender or ethnicity. Such demographic information is often used to correct for confounding, when alternative strategies are unavailable.

To determine whether the latent fitness representation contains information about patient age, we train a Bayesian ridge regression model to predict patient age from the representation ρ_i . We focus just on the $n = 505$ patients for which age information was available (non-missing). We hold out 12.5% of patients as test set. We repeat this process for the latent representations from each

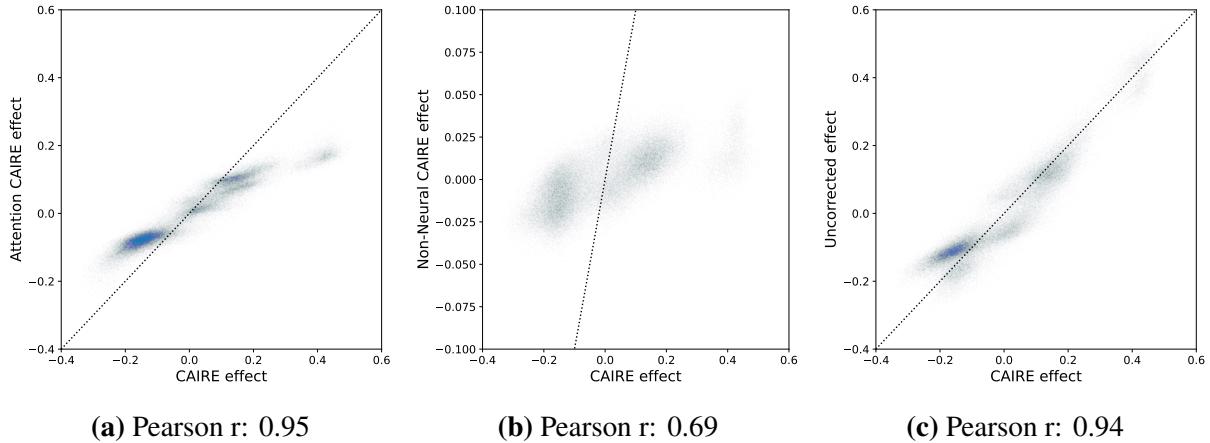


Figure S8: Comparison of COVID effect estimates from different methods. We compare effect estimates from CAIRE with those from Attention CAIRE (Figure S8a), Non-Neural CAIRE (Figure S8b) and the Uncorrected method (Figure S8c). Scatter plot shows 100,000 sequences randomly subsampled from the 11 million held out TCR sequences. The dotted black line shows the identity, where the two effects are equal.

of the ensemble of 24 CAIRE models, holding out a randomly chosen set of patients each time. On average over the ensemble, we find that the coefficient of determination (R^2) on the test set is 0.004, with a standard error of the mean of 0.009.

We next consider gender, training a logistic regression model to predict whether a patient is male or female. There were $n = 506$ patients with gender information available. The average accuracy over the ensemble of CAIRE models was 0.47 (standard error 0.01), while the average percentage of the test set that was female was 0.51.

Finally we consider ethnicity. We focus on predicting whether or not a patient is Caucasian, since other groups make up just a small percentage of the data (Appendix G.2). Using logistic regression, we find that the average accuracy over the ensemble was 0.84 (s.e. 0.01), while the average Caucasian fraction of the test set was 0.84.

Overall, then, we find no evidence to support the idea that the latent fitness representation reflects patient demographics. Note, however, that it also does not appear to the case that the latent fitness representation is arbitrary or purely random. First, it is predictive of the outcome variable (Table 3, Figure S7). Second, we computed the Euclidean distance matrix between patients' representations, that is $M_{ij} = \|\rho_i - \rho_j\|_2$ and found that it was stable across the ensemble of CAIRE models. In particular, to compare two models in the ensemble, we examined the Pearson correlation between the non-diagonal entries of their distance matrices, M . On average across pairs of models from the ensemble, the Pearson correlation was 0.70 (all p values for testing non-correlation were below floating point precision).

I.8 Binding prediction evaluation

To evaluate a model's ability to predict binding, we calculate the ROC AUC for separating binders from repertoire sequences, as described in Appendix H. We use MIRA binding data together with repertoire sequencing data from the same patients (Appendix G.5). We first consider separating sequences found to bind class I MHC epitopes from repertoire sequences drawn from the same set of patients (Table 4, column: Class I (all)). Then, we consider the same metric for class II MHC epitopes (Table 4, column: Class II (all)). We also repeat these calculations just for those sequences

that the model is confident has an effect ($\tilde{p} < 0.05$) (Table 4, columns: Class I and II (signif.)). We report the AUC standard error as $1/2\sqrt{\min(B, U)}$, where B is the number of binders and U is the number of repertoire sequences with unknown binding; note this is considered a somewhat conservative estimate of the true uncertainty [Cortes and Mohri, 2004].

For class I, we find that CAIRE’s estimates are predictive of binding. In particular, we find that sequences with more negative effects are more likely to bind an antigen (as indicated by a negative sign in Table 4). This is in line with the idea that sequences have negative effects because they give rise to overactive immune responses (Section 5.3): TCRs that bind more may be more likely to create an extreme immune response and damage healthy tissue.

For class II, we do not find clear evidence that CAIRE’s effect estimates are predictive of binding (Table 4), though note that less data is available for class II than class I (4,130 sequences bind class II, identified from binding experiments on 6 patients; 32,770 sequences bind class I, identified from binding experiments on 66 patients).

One possible concern is that these evaluation metrics pool data from many patients. Due to experimental variability, it may be the case that the distribution over sequences in the pooled set of repertoire sequences is slightly different from that in the pooled set of binders. As a robustness check, we also calculated the ROC AUC for discriminating binders (both class I and class II) from repertoire sequences *within* each patient, and then take the average ROC AUC across patients. The results confirm that CAIRE’s estimates are predictive of binding (Table 4, column: Patients (all)).

I.9 Other therapeutic approaches

We further considered the implications of CAIRE’s estimates for therapeutic approaches that deplete T cells with TCRs that bind a specific antigen [Moisini et al., 2008, Norville and Wood, 2023]. Here, a key design question is how to select an antigen that interacts with pathogenic TCRs. We therefore looked for antigens studied in the MIRA binding experiments that were enriched for TCRs with significant negative effects. We did not find any (binomial test, Benjamini-Hochberg adjusted p-value threshold of 0.05). This suggests that it is difficult to target and deplete the population of TCRs with negative effects, without depleting even more TCRs that have positive effects on clinical outcomes. To develop therapies that deplete TCRs with negative effects, it may be necessary to look outside the SARS-CoV-2 genome, for instance at human antigens.

Table 4: Binding prediction performance. ROC AUC values, with standard error. The sign of the prediction is in parenthesis: (-) indicates that sequences with more negative effects are more likely to bind, and (+) indicates that sequences with more positive effects are more likely to bind. NA values are reported in situations where no sequences have significant effects ($\tilde{p} < 0.05$). Note that for all class II values, 0.5 is within the 95% confidence interval (1.96 \times standard error).

	Class I (all)	Class I (signif.)	Class II (all)	Class II (signif.)
CAIRE	0.563 \pm 0.003 (-)	0.604 \pm 0.027 (-)	0.505 \pm 0.008 (-)	0.564 \pm 0.058 (+)
Attention CAIRE	0.568 \pm 0.003 (-)	0.526 \pm 0.109 (+)	0.507 \pm 0.008 (-)	0.651 \pm 0.500 (+)
Non-Neural CAIRE	0.566 \pm 0.003 (-)	NA	0.502 \pm 0.008 (+)	NA
Uncorrected	0.556 \pm 0.003 (-)	0.534 \pm 0.020 (-)	0.513 \pm 0.008 (-)	0.557 \pm 0.052 (-)
Patients (all)				
CAIRE	0.553 \pm 0.009 (-)			
Attention CAIRE	0.557 \pm 0.008 (-)			
Non-Neural CAIRE	0.556 \pm 0.008 (-)			
Uncorrected	0.549 \pm 0.007 (-)			

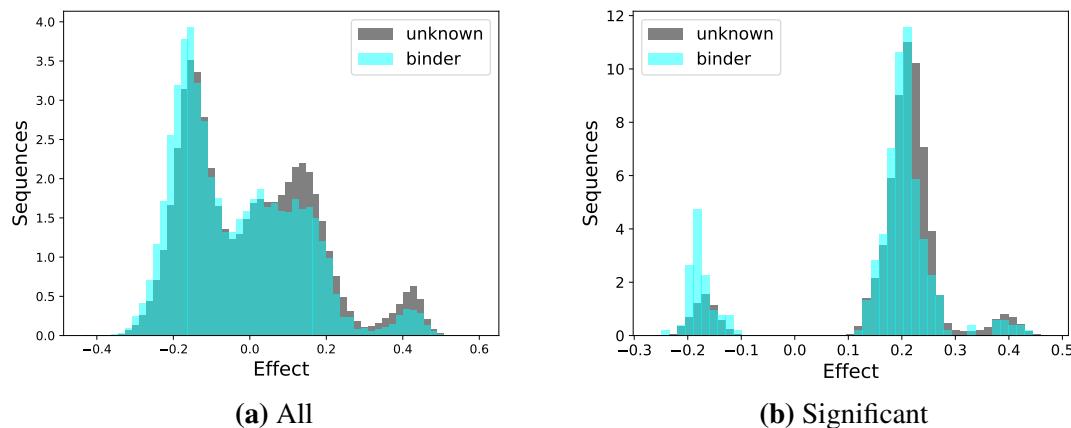


Figure S9: Effect distribution across binders. Distribution of estimated effects among sequences found to bind class I antigens in the MIRA assay (binders, blue) versus among general repertoire sequences, from the same patients the binders were found in (unknown binders, gray). (a) All sequences. (b) Only sequences with significant non-zero effects, $\tilde{p} < 0.05$.

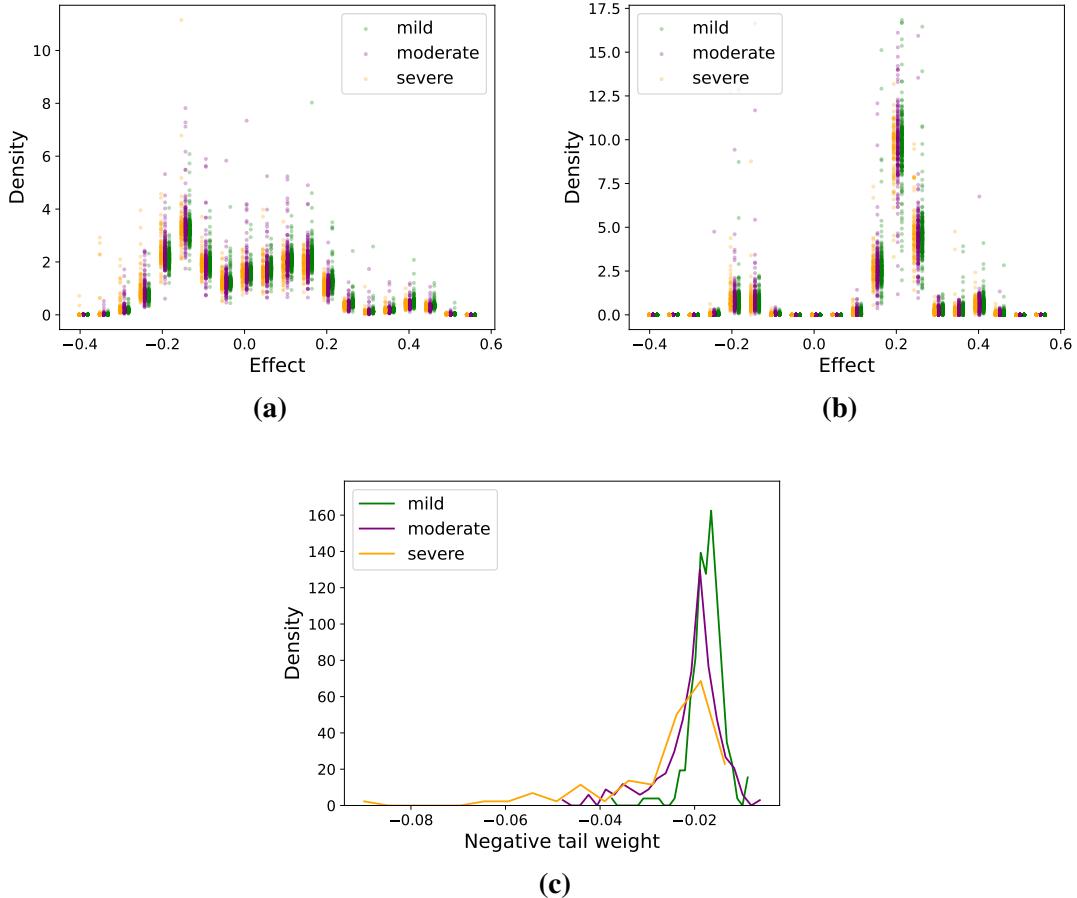


Figure S10: Distribution of effects across patients. (a) The distribution of TCR effects across individual patient repertoires. Each point at interval \mathcal{J} is an estimate of $\mathbb{P}_{A \sim q_i^a} [\text{ATE}(A; 0.1) \in \mathcal{J}]$ for a patient i . (b) The distribution of TCR effects among TCRs with significant effects, across individual patient repertoires. Each point is an estimate of $\mathbb{P}_{A \sim q_i^a} [\text{ATE}(A; 0.1) \in \mathcal{J} \mid \tilde{p} < 0.05]$ for a patient i . (c) Distribution of the burden of repertoire sequences with negative effects, across patients with different outcomes. Each point at interval \mathcal{J} is an estimate of $\mathbb{P}_{Q^a} [\mathbb{E}_{A \sim Q^a} [\text{ATE}(A; 0.1) \mathbb{I}(\text{ATE}(A; 0.1) < -0.2)]] \in \mathcal{J} \mid y]$ for an outcome $y \in \{-1, 0, +1\}$.

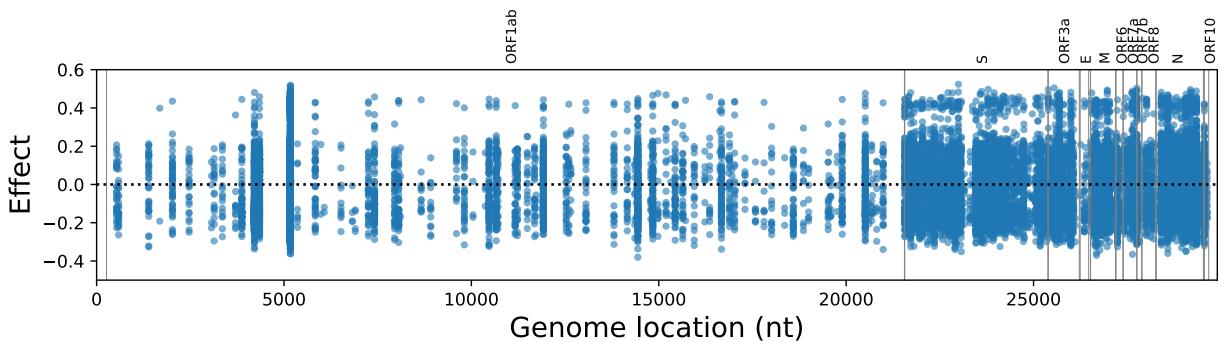


Figure S11: TCR effects across the SARS-CoV-2 genome. Same as Figure 5, except including all binders, not just those with significant effects.

Table 5: TCRs with the largest positive effects. Effect: the effect estimate from CAIRE, for $\epsilon = 0.1$. CDR3 β : the CDR3 β region of the TCR. Framework: the V and J gene for the TCR. Epitope(s): the SARS-CoV-2 epitope(s) the TCR was found to bind. Some entries have multiple epitopes because the experiment pooled them, so it is unknown precisely which of epitope the TCR binds. The table is divided into TCRs that were found to bind class I and class II MHC epitopes.

Effect	CDR3 β sequence	Framework	Epitope(s)
			Class I
0.436	CASTKEGRVATNEKLFF	V28-01+J01-04	HTTDPFLGRY
0.427	CASRRGQENEKLFF	V05-04+J01-04	TVLSFCAFA
0.426	CASSLATTGENEKLFF	V05-06+J01-04	VLSFCAFAV
0.414	CASCETHPVGYPNEKLFF	V27-01+J01-04	LLDDFVEII
0.400	CASSDRQGTNEKLFF	V27-01+J01-04	LLLDFVEI
0.397	CASSITGRANEKLFF	V19-01+J01-04	HTTDPFLGRY
0.396	CASSYRAGGNEKLFF	V06-05+J01-04	ILGTVSWNL
0.386	CAWKSEDRQGFNEKLFF	V30-01+J01-04	SNEKQEILGTVSW
0.383	CASSPNQQGTNEKLFF	V27-01+J01-04	LSPRWYFY
0.381	CASSDDQVGTANEKLFF	V18-01+J01-04	SPRWYFY
0.328	CASSLTGIEKLFF	V27-01+J01-04	YL
0.328	CASSQKTGGREKLFF	V04-01+J01-04	HTTDPFLGRY
			LLDDFVEII
			LLLDFVEI
Class II			
0.409	CASSQDQTGDNEKLFF	V03-01/02+J01-04	EDLKPRGQGVPI
			NTNNSP
			PNNTASWFTALTQHGKEDL
			QGVPI
			NTNNSPDDQIGYYR
			SSPDDQIGYYRRA
			TRRIRG
			TALTQHGKEDLKPRGQGV
0.409	CASSRTGGNEKLFF	V11-02+J01-04	ASFSTFKCYGVSP
			TKLN
			DLE
			GDEV
			RQIA
			PGQ
			YGV
			SP
			TKLN
			DLCFTNVYADSF
			VIRGDE
			PGQTG
			KIAD
			Y
			YADSF
			VIRGDE
			VRQIAPGQ
			YGV
			SP
			TKLN
			DLCFTNVYAD
			GRCDIK
			DLPE
			KEITV
			ATSV
			RTS
			LIAGHHL
			GRCDIK
			DLPE
			PKEITV
			ATSV
			RTS
			LYKLGA
			SRTS
			YKLGA
			ASQRVAGDS
			FLIVAAIV
			FITLC
			CFTLKR
			KTE
			SPKL
			FIRQEEV
			QELY
			SPIFL
			DFGGFNFS
			SQLPDP
			SKPSK
			DLL
			FNKVT
			LADAG
			FIKQYGDCLG
			DIAA
			PSKRSF
			IED
			LLFNKVT
			LAD
			QILPDPSK
			PSKRSF
			IED
			LL
			QYGDCLG
			DIAA
			RDLCIAQK

Table 6: Antigens that bind TCRs with strong positive effects. We examined the causal effect of the TCRs that bind each epitope. Focusing just on TCRs with a significant causal effect, we found that these antigens were specifically enriched for TCRs with beneficial effects on patient outcomes (binomial test, Benjamini-Hochberg adjusted p-value below 0.05). Mean effect sign: average sign of the estimated effect of TCRs that bind the epitope(s). Epitope(s): the specific SARS-CoV-2 antigen. Some entries have multiple overlapping epitopes because the experiment pooled these epitopes, so it is unknown precisely which of these epitopes the TCR binds. Note all discovered epitopes here are class I. Gene: the SARS-CoV-2 protein the epitope comes from (S: spike, N: nucleocapsid).

Mean effect sign	Epitope(s)	Gene
0.40	HTTDPNFLGRY	ORF1ab
1.00	APHGVVFL,APHGVVFLHV GVVFLHVTY,VVFLHVTYV	S
1.00	LSPRWYFYY,SPRWYFYYL	N
1.00	AYKTFPPTEPK,KTFPPTEPK	N
1.00	APSASAFFGM,AQFAPSASA ASAFFGMSR,SASAFFGMSR	N