Extrapolating missing antibody-virus measurements across serological studies

Highlights

- Large antibody-virus datasets can be described with a few simple patterns
- Matrix completion can exploit these patterns to predict unmeasured interactions
- Virus data can be extrapolated between studies
- Past results can guide future efforts by identifying the strongest/ weakest interactions

In brief

A central problem in immunology is to characterize how our antibodies inhibit the diverse pathogens we encounter in our lives. Here, we apply a well-studied technique called matrix completion that leverages patterns in partially observed antibody-virus inhibition data to infer unmeasured interactions. We predict tens of thousands of missing experiments in influenza and HIV-1 studies and quantify the expected error of our estimates. By harnessing matrix completion, future experiments could be designed that only collect a fraction of measurements, saving time and resources while maximizing the information gained.
Extrapolating missing antibody-virus measurements across serological studies

Tal Einav1,3,4,* and Brian Cleary2,3,*
1Basic Sciences Division and Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
2Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3These authors contributed equally
4Lead contact
*Correspondence: tal.einav@gmail.com (T.E.), bcleary@broadinstitute.org (B.C.)
https://doi.org/10.1016/j.cels.2022.06.001

SUMMARY

The development of new vaccines, as well as our understanding of key processes that shape viral evolution and host antibody repertoires, relies on measuring multiple antibody responses against large panels of viruses. Given the enormous diversity of circulating virus strains and antibody responses, comprehensively testing all antibody-virus interactions is infeasible. Even within individual studies with limited panels, exhaustive testing is not always performed, and there is no common framework for combining information across studies with partially overlapping panels, especially when the assay type or host species differ. Prior studies have demonstrated that antibody-virus interactions can be characterized in a vastly simpler and lower dimensional space, suggesting that relatively few measurements could predict unmeasured antibody-virus interactions. Here, we apply matrix completion to several large-scale influenza and HIV-1 studies. We explore how prediction accuracy evolves as the number of measurements changes and approximates the number of additional measurements necessary in several highly incomplete datasets (suggesting ~250,000 measurements could be saved). In addition, we show how the method can combine disparate datasets, even when the number of available measurements is below the theoretical limit that guarantees successful prediction. This approach can be readily generalized to other viruses or more broadly to other low-dimensional biological datasets.

INTRODUCTION

The constant evolution of viruses such as influenza and the human immunodeficiency virus (HIV-1) leads to viral escape and degraded immunity. As a result, the influenza vaccine must be periodically reformulated to focus our antibody response against currently circulating strains (Petrova and Russell, 2018). In the context of HIV-1, antibody cocktails are necessary to counteract the vast diversity of viral strains within each host (Wagh et al., 2018). Thus, the ability to characterize the spectrum of viral variants against panels of antibodies or antisera is critical to the development of new vaccine strategies and successful viral control.

Although thousands of new variants emerge each year, only a small fraction can be functionally characterized in terms of their effects upon the antibody repertoire. Assays quantifying binding, neutralization, or hemagglutination inhibition (HAI) are time and resource intensive, necessitating small antibody-virus panels that rarely exceed several dozen strains (Wrammert et al., 2011; Li et al., 2012; Georgiev et al., 2013; Kong et al., 2015; Creanga et al., 2021). Crucially, virus panels often differ between studies, making it difficult to translate the lessons learned in one context to predict new viral behavior.

Nonetheless, such small-scale studies have helped shed light on how viral infection or vaccination reshape the antibody response (Amanna et al., 2007; Thompson et al., 2016; Lee et al., 2019a; Arevalo et al., 2020) as well as how the existing antibody repertoire constrains viral evolution (Perelson et al., 2012; Dingens et al., 2019; Lee et al., 2019b). Moreover, an increasing number of large-scale studies have probed how our antibody response copes with the diverse array of viruses we encounter. For example, each year the influenza surveillance network surveys ~100,000 viruses against ~10 reference sera to determine whether an antigenically distinct virus strain has emerged. Combining this surveillance data with insights from small-scale studies could increase the resolution of antibody-virus interactions and lead to improved vaccine selection (Petrova and Russell, 2018).

This ability to combine datasets is especially feasible in the context of antibody-virus interactions, where past studies have repeatedly shown that the data are low-dimensional (Lapedes and Farber, 2001; Ndifon, 2011). Strains from the same lineage often elicit similar antibody responses and can be grouped together into antigenic clusters (Smith et al., 2004). Different antibodies targeting the same epitopes, or sera from individuals with similar exposure histories, can show correlated activity across viral strains (Fonville et al., 2014; Kucharski et al., 2015). The profile of serum inhibition from an individual exposed to two antigenically distinct viruses can be expressed as a
combination of profiles from two individuals, each exposed to one of the viruses (Carter et al., 2016). Mathematically, these features are consistent with a low-rank structure (Figure S1).

The theory of compressed sensing and the related field of matrix completion describe how such structures can be leveraged in the setting of under-sampled experiments (Candes and Recht, 2009; Cleary and Regev, 2020). These techniques have been extensively studied theoretically and applied in many settings from online recommender systems to magnetic resonance imaging (Candes and Tao, 2010). With matrix completion, the goal is to use a relatively small number of observations (individual matrix entries; here, corresponding to HAI or neutralization measurements between a virus-antibody or virus-serum pair) to identify low-rank features that can be used to infer missing values. Biologically, such inferences are possible because antibodies cross-react and often exhibit similar behavior against similar viruses. Although the composition of each serum is unknown and may be highly varied, the low-dimensional nature of these interactions suggests that a serum’s measurements against a few viruses can predict its behavior against many other strains.

In the context of antigenic characterization, matrix completion has been used to refine the calculation of antigenic distance between strains (Cai et al., 2010; Ndifon, 2011). However, prior work has not sufficiently addressed questions of sample complexity (describing how the magnitude of errors will change based on the number of available measurements), which is critical for the successful design of under-sampled experiments. Moreover, although earlier studies have performed matrix completion on the data within a given study (what we will call intra-table matrix completion, Figure 1A), the limits of inter-table matrix completion are not clear (Figure 1B). It remains to be demonstrated to what extent partially overlapping studies conducted in varying geographic regions and points in time can be integrated, whether different types of studies (e.g., infection versus vaccination or using fluorescence versus HAI; Fonville et al., 2014; Vinh et al., 2021) can be combined, and whether studies across species (particularly human versus ferret in the case of influenza) can successfully inform each other in the context of matrix completion. Successful inference in these tasks could vastly increase the amount of data used to study viral evolution or develop vaccination strategies and could be applied to improve ongoing experimental design.

Here, we develop a low-rank matrix completion framework that can analyze the increasing number of serological studies of the antibody response against viruses. We focus on three of the largest such datasets: (1) the Fonville influenza dataset consists of six studies (1 ferret study, 1 human infection study, and 4 human vaccination studies; collectively 1,147 sera × 81 viruses), all measured using HAI (Fonville et al., 2014). (2) The Vinh influenza dataset contains one study (human infection study, 24,000 sera × 6 viruses) also measured using HAI (Vinh et al., 2021). (3) The Catnap HIV-1 dataset contains two categories of data (monoclonal antibody data, 373 sera × 933 viruses; polyclonal serum data, 40 sera × 71 viruses) combining antibody-virus neutralization measurements from over one hundred studies (Yoon et al., 2015).

Such experiments commonly contain a subset of missing measurements. Ideally, subsequent analyses can ignore these missing values and use all available data—however, in some cases, such as quantifying the dimensionality of the antibody response using singular value decomposition, entire rows, or columns must be removed to obtain a complete subset of data (Figure 1A, red lines). Moreover, measurements across multiple studies are difficult to combine since virus panels tend to only partially overlap, further constraining the number of measurements (Figure 1B). For example, only 9/81 viruses were included in all six Fonville studies and 0/933 of the Catnap viruses were included across the full database.
In this work, rather than trimming down to a complete subset of data, we instead harness the low dimensionality of the immune response and extrapolate missing antibody-virus measurements both within and between studies. Mathematically, missing values are imputed by minimizing the nuclear norm of the antibody-virus data to create a complete and low-rank matrix. This framework can be implemented with fewer than 10 lines of code in Python or Mathematica (see STAR Methods and the linked GitHub repository). In the following sections, we illustrate this approach using progressively more complex forms of matrix completion, creating a framework that naturally incorporates antibody-virus measurements from multiple studies to inform future efforts. In the final section, we show how datasets can be combined to predict hundreds of thousands of new measurements.

RESULTS

Antibody-virus data are consistent with a low-rank structure

In matrix completion, a partially observed matrix $M \in \mathbb{R}^{n_1 \times n_2}$ is described using a low rank $r$ (with $r < \min(n_1, n_2)$), so that the interaction between antibody (or serum) $i$ and virus $j$ is given by $\sum_{k=1}^{r} u_{ik} v_{kj}$ (Figure S1). This dramatically simplifies $M$, representing it with $r(n_1+n_2)$ entries rather than the full $n_1 n_2$. The goal of matrix completion is to find the feature set $u_{ik}$ for antibodies and $v_{kj}$ for viruses that best approximates $M$, while simultaneously minimizing $r$ to describe the data as simply as possible (see STAR Methods for more details).

We first quantified how well a low-rank approximation can characterize the influenza and HIV-1 datasets. Every dataset we evaluated is considerably low-rank in comparison with its naive dimensionality, with ~95% of variance in entries explained by rank 6–23 approximations, depending on the dataset (Table S1). These low-rank approximations minimally shifted the existing measurements, with a root-mean-squared error (RMSE) $\leq 0.06$ for the log$_{10}$ titers across all studies (i.e., unlogged titers are shifted by $10^{0.06} = 1.1$-fold on average, representing near perfect fidelity).

These results hold for the nearly complete influenza studies (Fonville Studies 5–6; each 160 sera $\times$ 20 viruses and rank 8–9) as well as the more incomplete studies (Fonville Studies 1–4; between 35–324 sera $\times$ 57–74 viruses and rank 6–8) (Table S1; Figure S2). Interestingly, we observed that the low rank of the Fonville data is similar to the number of antigenic clusters ($n = 14$) in the combined panel of 81 viruses (Fonville et al., 2014), although substantial differences between viruses in the same antigenic cluster are observed in the completed matrix (Figure S3). We hypothesize that when a new antigenic cluster emerges and the diversity of immunodominant epitopes increases, viruses in that cluster will exhibit distinct phenotypes that will increase matrix rank. The low-rank result also holds for large tables with hundreds of antibodies and viruses, including the monoclonal antibody Catnap database (373 antibodies $\times$ 933 viruses and rank 23). The larger rank in the Catnap dataset reflects a greater number of viruses (~12x more than the next largest study) in addition to the larger diversity between the viruses in HIV-1 versus influenza data.

This consistent, low-rank structure suggests that substantial information about missing entries may be contained in the already measured values. According to established theoretical results, if we wish to complete the missing entries of an $n_1 \times n_2$ matrix with minimal error, the number of available measurements ($m$) needs to be proportional to the rank ($r$) of the complete matrix and proportional to $n \log(n)$ (with $n = \max(n_1, n_2)$) (Candes and Tao, 2010; Recht, 2011). The number of necessary measurements is also proportional to the coherence of the low-rank projection ($\mu$), which quantifies how concentrated the low-rank factors are on individual entries. Matrices with larger coherent—which in this context corresponds to individual antibodies or viruses with unique behavior that is orthogonal to the behavior of all other antibodies or viruses—require more measurements for successful recovery. Together, at least

$$m > C \mu r n \log(n)$$

(Equation 1)

measurements are necessary, for some positive constant $C$ that is context specific (we approximate $C$ below) (Candes and Tao, 2010).

Although these results establish the sample complexity necessary for recovery with minimal error, there may be practical scenarios that deviate from some underlying assumptions of the theory. For example, the theoretical result assumes the available entries have been sampled uniformly at random from the matrix, but in reality, available data are likely to have been sampled in a biased (non-uniform) fashion and may require more measurements than the theoretical limit suggests. However, even when the number of samples is below this theoretical limit, the recovered values may nonetheless contain useful information about missing entries that might be used to effectively prioritize future experiments. Below, we investigate how the error in completed entries evolves under increasing sample complexity and under uniform and non-uniform sampling.

Sample complexity under uniform random sampling

To quantify the accuracy of matrix completion within a single study using uniform random sampling (which we call intra-table completion), we split the available data in each study into observed measurements and a withheld validation set, and varied the fraction of entries withheld while completing the matrix using the observed entries as input (Figure 1C). As the observed fraction approaches 100% of available entries, we approximate how accurately the truly missing values in the dataset can be recovered (shown by the purple boxes for the schematic curves in Figure 1D).

In the context of influenza, we first withheld 50% of measurements in Fonville Study 1 (the first of six studies within Fonville et al., 2014) and used matrix completion to impute the values of all missing and withheld titers (Figure 2A, withhold data [gold] are drawn from the available measurements [grayscale in Figure S2]). We then quantified the accuracy of the withheld validation set, which exhibited a Pearson’s correlation coefficient of $r^2 = 0.8$ (i.e., 80% variance explained) and an RMSE $= 0.4$ for the log$_{10}$ titers (Figure 2B). To assess the theoretical limit for accurate recovery from Equation 1 (carried out below), we estimate the complexity of the completed data as the rank required to explain 95% of the variance, with rank $= 7$ for the data shown
This analysis is repeated for all six Fonville studies at varying fractions of observed data, with the accuracy of each completion quantified using Pearson’s correlation coefficient ($r^2$) and the RMSE of the log10(titers) averaged over 10 runs to sample different subsets of data.

For nearly complete studies with few missing values, we expect the final estimated error to approach the experimental accuracy of the assay, whereas the error for studies with a sizable fraction of missing values will depend on how the number of measurements compares to the bound discussed above (Equation 1). In the nearly complete tables (Fonville Studies 5 and 6), we indeed observe that the RMSE approaches 0.3 (in log10 units, Figure 2D), comparable with the $10^{0.3} = 2$-fold accuracy of the HAI assay (see Figure S8 of Fonville et al., 2014). As more entries are observed, the RMSE decreases, as expected (Figure 2D). The same trend is observed in the more incomplete influenza datasets, as well as the HIV-1 Catnap database (Figure 3), with the final estimated RMSE for the log10 titers across all tables varying from 0.2 in Fonville Studies 3–6 (90% of the full dataset available) to 0.8 in Catnap (13% available). Therefore, by varying the fraction of data observed, we can estimate both the error of the imputed titers and whether additional measurements will lead to markedly better predictions. We found these results, as well as those presented below, to be robust across a wide range of values for the sole hyperparameter $\lambda$ (STAR Methods).

Although Studies 1, 2–4, and 5–6 gave rise to different RMSE curves, the responses are highly similar when we divide the number of observed measurements by the right-hand side of Equation 1 (Figure S4). Thus, as predicted by previous theoretical results (Candes and Tao, 2010), there is a universality describing when matrix completion will start to fail based upon the size, rank, and coherence of each dataset.

Since all six Fonville studies impute the missing entries with an RMSE on the order of the experimental error, we can use these curves to determine how fewer measurements could produce equally successful completions. In brief, because the experimental error may vary by dataset and assay, we determine when the slope of each RMSE curve bottoms-out (slope $\approx -0.5$, see STAR Methods). We find that in Fonville Study 1, 67% of measurements were needed (implying $\lambda \approx 1,000$ measurements could have been “saved”); in Fonville Study 2, 37% were needed ($\lambda \approx 12,000$ saved measurements); in Studies 3 and 4, 40% were needed ($\lambda \approx 20,000$ saved measurements); and in Studies 5 and 6, 55% were needed ($\lambda \approx 3,000$ saved measurements) (inset of Figure 2D). The total across all studies, compared with measuring every value (including those missing in the published datasets), is $\lambda \approx 35,000$ potential saved measurements of a total of $\lambda \approx 60,000$ possible measurements (58%).

We next examined intra-table completion in the context of HIV-1, where we used the more incomplete Catnap database to approximate the number of additional measurements

---

**Figure 2.** Completing antibody-virus measurements within influenza studies

(A–C) Low-rank matrix completion of Fonville Study 1 using 50% of all antibody-virus interactions. (A) Using available measurements (titers in grayscale) to infer the missing values (blue) and the withheld validation set (gold). The top panel shows the input matrix and the bottom panel the completed matrix. (B) Predicted versus measured titers for the validation set withheld from the input matrix. (C) The variance of the completed matrix explained with a rank $r$ approximation. The rank of a study is defined as the value $r$ necessary to explain 95% of the variance (rank = 7 for the data shown). (D) Ten iterations of intra-table completion at different fractions of observed data were performed for all six Fonville studies (individual lines), with the (mean ± standard deviation) Pearson’s $r^2$ and RMSE shown. (A–C) represent 1 of 10 iterations at an observed fraction of 50% (boxed points in D). The results for Studies 2–4 and Studies 5–6 are nearly overlapping. The markers on the x axis of the RMSE plot denote the total fraction of antibody-virus interactions measured in each study (Table S1). Studies 1–6 contain 2,200, 12,000, 14,000, 16,000, 3,200, and 3,200 measurements, respectively (Figure S2). Inset: the fraction of data in each study that is required for accurate matrix completion (gray), excess measurements that could have been inferred (orange), or missing measurements (blue).
necessary to achieve the lower bound RMSE. In monoclonal antibody studies, matrix completion using all available data led to an $r^2 = 0.8$ and RMSE = 0.8 for the missing values, with both metrics expected to rapidly improve with additional measurements (Figure 3). To approximate the minimum number of additional measurements necessary to achieve the lowest possible RMSE, we first use the available data in each Fonville study to estimate the coherence $\mu$ from Equation 1 (see STAR Methods). We then use the number of necessary measurements empirically determined above, together with the theoretical lower bound in Equation 1, to find the value of the constant, $C$, in each Fonville study (Table S2) and average this value across the studies. Finally, given this constant, the size of the monoclonal antibody Catnap table, and estimates of its rank and coherence, we approximate the number of additional measurements necessary for successful completion (Table S2). This suggests that 27% of all possible measurements are necessary, implying that ~48,000 of the ~300,000 missing antibody-virus pairs need to be measured. We repeated this analysis for the polyclonal antibody Catnap data, which, like the Fonville tables, was sufficiently complete to get an empirical estimate of the fraction of measurements necessary for successful completion (Table S2). This form of completion is especially useful for rapidly evolving pathogens such as influenza or HIV-1, where studies published in the same year will often only have partially overlapping virus panels (Figure 1B). Relatedly, when predicting the results of experiments done in coming years using previously generated data, which we can evaluate in the Catnap database using the timestamps of each measurement, the available or missing data might be biased toward a few antibodies or viruses.

We first asked if inter-table completion can predict the inhibition of viruses that were not included in a given study. This form of completion is especially useful for rapidly evolving pathogens such as influenza or HIV-1, where studies published in the same year will often only have partially overlapping virus panels (Figure 1B). Relatedly, when predicting the results of experiments done in coming years using previously generated data, which we can evaluate in the Catnap database using the timestamps of each measurement, the available or missing data might be biased toward a few antibodies or viruses.

Non-uniform sampling: Withholding a virus via inter-table completion

In the previous section, we examined the data in each study individually (intra-table completion), and the withheld validation set was uniformly sampled from all available measurements. In practice, missing or available measurements may not be spread uniformly across the matrix. For example, when combining data from different studies, which we call inter-table completion, missing measurements might appear in a more structured way (e.g., as blocks of missing values in Figure 1B). We then use the number of necessary measurements empirically determined above, together with the theoretical lower bound in Equation 1, to find the value of the constant, $C$, in each Fonville study (Table S2) and average this value across the studies. Finally, given this constant, the size of the monoclonal antibody Catnap table, and estimates of its rank and coherence, we approximate the number of additional measurements necessary for successful completion (Table S2). This suggests that 27% of all possible measurements are necessary, implying that ~48,000 of the ~300,000 missing antibody-virus pairs need to be measured. We repeated this analysis for the polyclonal antibody Catnap data, which, like the Fonville tables, was sufficiently complete to get an empirical estimate of the fraction of measurements necessary for successful completion, and found good agreement between the empirical and theoretical results (44% and 41%, respectively; Table S2). We stress, however, that the theoretical estimates are a rough approximation, meant more to identify the scale for potential under-sampling, rather than attempting to define a specific point at which measurements should stop.

Non-uniform sampling: Withholding a virus via inter-table completion

In the previous section, we examined the data in each study individually (intra-table completion), and the withheld validation set was uniformly sampled from all available measurements. In practice, missing or available measurements may not be spread uniformly across the matrix. For example, when combining data from different studies, which we call inter-table completion, missing measurements might appear in a more structured way (e.g., as blocks of missing values in Figure 1B). Relatedly, when predicting the results of experiments done in coming years using previously generated data, which we can evaluate in the Catnap database using the timestamps of each measurement, the available or missing data might be biased toward a few antibodies or viruses.

We first asked if inter-table completion can predict the inhibition of viruses that were not included in a given study. This form of completion is especially useful for rapidly evolving pathogens such as influenza or HIV-1, where studies published in the same year will often only have partially overlapping virus panels (Figure 1B). Relatedly, when predicting the results of experiments done in coming years using previously generated data, which we can evaluate in the Catnap database using the timestamps of each measurement, the available or missing data might be biased toward a few antibodies or viruses.
Since every virus had available measurements in at least two studies, the withheld virus was still functionally characterized, albeit in the context of a different study. We found that on average, the imputed titers had an RMSE = 0.3 ± 0.2 (Figure 4B, gray bands), corresponding to roughly 2-fold error when predicting the behavior of a virus that is entirely absent from a given study. Moreover, we examined whether prediction accuracy depends on the “distance” between the withheld virus and the remaining panel (based on the average fold difference of measured titers across all sera). Surprisingly, we found no correlation between RMSE and virus distance, demonstrating that even viruses that have no close homolog in the virus panel can be accurately predicted.

Although the majority of inter-table completions led to accurate predictions, some virus-study pairs showed a large RMSE (Figure 4B). Inter-table completion performed best for Fonville studies #3–4 and #5–6. Both pairs of studies used the same virus panel, and each contained human vaccination data collected in consecutive years. Therefore, these studies represent the ideal scenario to apply inter-table completion, since virus trends in one study should be highly informative of the trends in the other study. Indeed, removing any virus from these four studies led to low reconstruction error (RMSE = 0.2 ± 0.1 for Studies 3–4 and RMSE = 0.3 ± 0.1 for Studies 5–6). To quantify how much effort could have been saved in each pair of studies, we completely removed multiple viruses and computed the average RMSE for the imputed values (Figure S6). In each case, removing half the viruses from one study only increased the average RMSE by 30%. This suggests that when two contemporary studies with similar experimental designs have partially overlapping virus
panels, the two studies can be merged and accurately matrix completed.

On the other hand, Fonville Study 1 (ferret study) and Study 2 (human infection study) did not have a paired study with an identical virus panel and experimental design (Figure S7). Hence, these two studies represent a harder but more common scenario for inter-table completion. Although many of the viruses in these first two studies had accurate inter-table completions (average RMSE = 0.5 ± 0.2 for both studies), there were some extreme outliers. For example, the hardest virus to predict (A/Auckland/20/2003) appeared in Studies 1, 3, and 4; it exhibited a large RMSE = 1.2 when removed from Study 1 but a small RMSE = 0.3 when excluded from Study 3 or 4. Moreover, restricting the inter-table completions to a subset of studies—such as imputing human infection data (Study 2) using human vaccination data (Studies 3–6) without ferret data (Study 1)—leads to marginally worse completions, and more generally, including additional and varied datasets improves predictions (Figure S7). Indeed, when combining all Fonville studies, 99/131 = 75% of the inter-table completions for viruses in Study 1 or 2 had an RMSE < 0.6, demonstrating that behavior of many viruses can be inferred across disparate studies. These results suggest that inter-table completion should improve as more datasets are combined, providing a data-driven framework to unify diverse datasets and extend their predictions. Moreover, the following sections demonstrate that inter-table completion can expedite experiments even in the low-data limit where the imputed values are less accurate.

Non-uniform sampling: Predicting between monoclonal and polyclonal antibody data

We next evaluated a different form of inter-table completion using monoclonal antibody data to impute polyclonal serum measurements (and vice versa) in the Catnap database. Although both studies measure the amount of antibodies required to neutralize a virus by 50%, monoclonal antibodies represent a single, defined antibody, whereas polyclonal serum contain an unknown mixture of antibodies. Hence monoclonal antibodies are measured in terms of the 50% inhibitory concentration (IC50 in μg/mL units; lower values represent a more potent antibody), whereas sera are measured by the dilution necessary to achieve 50% neutralization (ID50 in dilution units; higher values represent a more potent serum). This more difficult case of inter-table completion must combine these disparate types of data as well as compensate for greater incompleteness of the data, with a far larger number of monoclonal antibody measurements (373 antibodies × 933 viruses) than polyclonal data (40 sera × 71 viruses; each of these viruses is also present in the monoclonal data).

Consistent with these expectations, we observed higher inter-table RMSEs than in Fonville studies where all the data were of the same type. The average error when withholding viruses from the monoclonal study was RMSE = 1.2 ± 0.3, whereas the average error when withholding from the polyclonal study was RMSE = 0.7 ± 0.3, implying an absolute prediction error of roughly 16- or 5-fold, respectively. However, given that the correlation between the imputed and missing entries was also somewhat high (average ρ2 of 0.47 and 0.49 when withholding from the monoclonal and polyclonal studies, respectively), the imputed data may contain useful information. For example, these predictions could differentiate between the weakest and strongest of the missing interactions.

Hence, despite the larger magnitude of errors, we hypothesized that inter-table completion could identify the most potent antibodies or sera. For each virus-study pair in the merged monoclonal-polyclonal Catnap data, we sorted the inferred entries for the withheld virus from most neutralizing to least neutralizing. We then asked whether prioritizing future experiments in this order would identify highly inhibitory antibodies (IC50 < 0.1 μg/mL for monoclonal; ID50 > 1,000 dilutions for polyclonal data) more effectively than performing future experiments in a random order. Averaging across viruses withheld from the monoclonal study, only the first 36% of prioritized experiments were necessary to identify 80% of all highly inhibitory antibodies (Figure 4C). Similarly, the top 30% of prioritized experiments were necessary to identify 80% of the top hits in the polyclonal study (Figure 4C). Both cases improve upon a random search, which would require 80% of measurements to be carried out to find 80% of the most inhibitory interactions. Moreover, these results hold even when we exclude antibodies and sera with similar neutralization profiles (Figure S8). Thus, even when matrix completion predictions have large errors, they nonetheless contain actionable information for the efficient design of future experiments.

The surprising fact that polyclonal serum measurements can inform monoclonal antibody data (and vice versa) suggests that the wealth of existing datasets in each category should be “synergistically combined” to expedite future experiments. For example, when a new influenza virus emerges, it is often rapidly characterized against ferret sera, and one immediate goal is to understand whether monoclonal antibody therapies will still be effective against this new variant (Figure 4D, bottom-left). Matrix completion would identify which of the antibodies would be the most potent (or weakest) against the new virus to directly address this question. Alternatively, if a new virus is measured against existing monoclonal antibodies, matrix completion would determine whether individuals with the strongest protection against previous strains will still have strong protection against the new virus (Figure 4D, bottom-right).

Non-uniform sampling: Using timestamps to predict measurements across time

We next examined how well existing measurements can predict future experiments. To that end, we turned to the dozens of studies comprising the monoclonal antibody-virus interactions in the Catnap database that together create the largest matrix we analyzed (350,000 interactions; Figure 3A). Each measurement is timestamped with the year it was deposited, enabling us to quantify how well past measurements could predict future data.

As noted from our intra-table completion analysis above, 87% of entries are missing from this monoclonal antibody data, and the imputed titers exhibited a relatively high estimated RMSE = 0.8 (signifying 6-fold error). This large error not only arises from the large fraction of missing entries but also from the considerable divergence between the HIV-1 viruses in the Catnap database, which differ by 200 amino acids on average, far exceeding the average 30-amino acid difference between the Fonville influenza viruses. It is easier to matrix complete a virus’s titers when a similar virus exists in the dataset, and conversely, the accuracy of matrix completion decreases as the distance to the nearest
We used the timestamp associated with each Catnap measurement and considered data taken before a given year (e.g., in 2017 or before) to predict the missing values that were subsequently measured (e.g., in 2018 or later; Figure 5A). In doing so, we only consider the antibodies and viruses measured both before 2017 and after 2018 (in the example of Figure 5A, we ignore the right-most column and the bottom row). We then ordered the predictions from strongest to weakest (Figure 5B) and determined the number of strong interactions (defined as an IC_{50} ≤ 0.1 μg/mL) identified by N experiments.

When using matrix completion to predict the most potent antibody-virus interactions, only 230/1,230 = 20% of measurements were required to identify 160/200 = 80% of the strong interactions found in 2018 or later (Figure 5C). In contrast, searching through the antibody-virus pairs at random would have required 990 measurements on average before 80% of the strong interactions could have been found, necessitating 4x more work. A similar trend was found when searching for the weakest antibody-virus interactions or when analyzing the Catnap data before/after a different year (Figure S10). In this manner, matrix completion can sift through the wealth of available data to find new antibody-virus interactions with especially high or low titers.

**Applying inter-table completion to extend a small virus panel measured against 24,000 sera**

Finally, we applied inter-table completion to a recent serological study to generate hundreds of thousands of predictions that can guide future experiments. Specifically, we combine measurements from the Fonville 2014 influenza study described above (1,147 sera × 81 viruses) with another large-scale influenza study by Vinh et al. (24,000 sera × 6 viruses; all six viruses have equivalent strains in the Fonville panel, as described in the STAR Methods) (Fonville et al., 2014; Vinh et al., 2021). For each virus in the Fonville panel whose behavior can be extrapolated to the Vinh sera, we generate 24,000 data points.

To ascertain which viruses can be predicted between these two studies, we first analyzed the 6 viruses in the Vinh study and repeated our earlier procedure, withholding measurements in the Vinh study for one virus at a time (Figure 6A). For this process, we used the completed Fonville data to fill in missing measurements against these six viruses, thereby leveraging all Fonville sera to inform the Vinh predictions. The predicted titers of the four viruses isolated between 2005 and 2011 exhibited a strong correlation (r^2 ≥ 0.8) with the withheld values and could effectively identify the potent antibody responses (Figures 6B and S11). For all four strains, ≥80% of potent interactions would have been discovered by testing only 20% of sera sorted by matrix completion. In contrast, the two viruses isolated in 2003 and 1968 exhibited a weaker correlation and worse discovery rate (Figures 6B and S11). Therefore, we restricted our application of inter-table completions to the 17 Fonville viruses isolated between 2005 and 2011. We combined the Vinh and Fonville measurements of the 6 viruses in the Vinh panel, together with Fonville measurements of one of the 17 additional viruses, and predicted the measurements for this additional virus against the 24,000 Vinh sera (Figure 6A, blue squares). We visualize the completed matrix by highlighting the 20% most potent sera for each of the Vinh and Fonville viruses (colored or black bars in Figure 6C; white bars represent the remaining 80% of sera).
As a sanity check, we find that 3,000 sera that were uniformly potent against the 4 Vinh viruses isolated between 2005 and 2011 were also predicted to be highly potent against the Fonville viruses from this same time period (Figure 6C, cyan box). Very few of these sera are predicted to inhibit 0–1 of the Fonville viruses (10 sera), and the majority of sera are predicted to inhibit ≥10 Fonville viruses (2,700 sera).

The expanded virus panel further enables us to detect inhibition patterns that cannot be discerned from the Vinh data alone. For example, a set of 300 sera were outside the top 20% of potent interactions for all six Vinh viruses; however, they were in the top 20% for at least one Fonville strain (Figure 6C, pink box). These sera exhibited a wide range of distinct inhibition profiles, ranging from potently inhibiting just 1 of the Fonville viruses (70 sera) to ≥10 viruses (80 sera). In this way, sera with specific inhibition profiles of interest against the extended virus panel—from the exceptionally broad to especially narrow—can be picked out from the 24,000 sera for further study.

DISCUSSION

The low dimensionality of antibody-virus data provides the mathematical foundation to fill in missing entries in a partially observed table of binding, neutralization, or HAI measurements.

Figure 6. Predicting an expanded virus panel using inter-table completion

(A) Combining data from the two influenza studies (Vinh et al., 2021) (24,000 sera × 6 viruses) and (Fonville et al., 2014) (1,147 sera × 81 viruses). Six of the Fonville viruses are equivalent to the Vinh strains, whereas 17 other Fonville viruses were isolated between 2005 and 2011 and used to extrapolate the Vinh responses (STAR Methods). (Left) Each of the six viruses is individually withheld and predicted via matrix completion. (Right) The complete Fonville and Vinh data predict how these 24,000 sera would interact with one of the 17 additional Fonville viruses.

(B) When one of the Vinh viruses is withheld (gold column), its imputed ID50s are ordered from the strongest to weakest. After carrying out a fraction of these prioritized experiments (x axis), the number of strong antibody-virus interactions discovered (with ID50 > 1,000) is shown for each virus.

(C) For each virus, the 20% of sera with the strongest predicted interactions are shown as either colored or black bars (white bars represent the 80% of weaker interactions). The sera are first ordered by their interaction against Virus 1 (potent sera shown at the bottom-left in blue), then Virus 2 (the bottom gold segment shows sera potent against Virus 1 and 2; the short gold segment above that denotes sera only potent against Virus 2), then Virus 3, and so on.
This same principle underlies the field of compressed sensing, where a sparse signal can reconstruct a complete image, enabling dramatically faster data acquisition. In the context of immunology, low dimensionality enables us to readily combine existing studies to infer missing measurements with high accuracy or, in cases of severe data limitation, rationally prioritize a minimal set of experiments to identify important antibody-virus interactions. With completed data, sera across studies can be compared and correlated against the same set of viruses, combining data instead of throwing it away (Figure S3). Given the increasing number of large-scale serological studies, the power of data-driven techniques such as matrix completion will continue to grow, offering a simple framework that can unify diverse datasets and vastly increase the amount of available data.

We suggest several ways to harness matrix completion and facilitate future serological studies. First, quantifying how a large panel of antibody responses inhibits a large number of viruses does not require measuring each interaction. If the experiment can be performed in several rounds (e.g., randomly selecting 20% of the unmeasured antibody-virus pairs each time), then the error of imputing the missing values can be calculated after each round (Figure 2D; Cao et al., 2018). When this error becomes comparable with the error of the assay (typically 2-fold for HAI or neutralization measurements) or when the error bottoms-out, the remaining values can be filled in via matrix completion. In the context of the Fonville 2014 studies, this amounts to saving tens of thousands of measurements (corresponding to 50% fewer experiments).

Second, when only a small number of viruses or sera will be screened (Harvey et al., 2016; Vinh et al., 2021), the few viruses or sera chosen should be as functionally orthogonal to one another as possible to maximize the information gained. A natural extension is to use matrix completion to determine a “minimal covering,” the smallest set of viruses needed to predict a full virus panel. Looking ahead, we hypothesize that the accuracy of matrix completion could define virus-virus distance (with poorly covering,’’ the smallest set of viruses needed to predict a full viral extension is to use matrix completion to determine a “minimal another as possible to maximize the information gained. A natural extension is to use matrix completion to determine a “minimal covering,” the smallest set of viruses needed to predict a full virus panel. Looking ahead, we hypothesize that the accuracy of matrix completion could define virus-virus distance (with poorly covering,’’ the smallest set of viruses needed to predict a full viral extension is to use matrix completion to determine a “minimal another as possible to maximize the information gained. A natural extension is to use matrix completion to determine a “minimal covering,” the smallest set of viruses needed to predict a full virus panel. Unlike with intra-table completions, where a fraction of the data can be withheld to quantify the expected error of the imputed values, there is currently no analogous method to quantify the expected error from inter-table completion. Hence, a fraction of interactions needs to be measured to verify the results of such matrix completions.

Even in cases where the error of imputed values is unacceptably large (for example, if the majority of measurements are missing or if the diversity of antibodies/viruses results in large coherence), inter-table completion can nevertheless pinpoint the strongest or weakest interactions that are often of greatest interest. For example, although only 16% of all 1,230 HIV-1 antibody-virus measurements from 2018 to 2020 had a strong IC50 ≤ 0.1 μg/mL, matrix completing all data available in 2017 determined a minimal set of 230 experiments that identified 80% of these strong IC50 values (Figure 5C).

One surprising feature of our analysis was that polyclonal sera could identify the strongest neutralization measurements for monoclonal antibodies (and vice versa) in the context of HIV-1, although the composition of antibodies within these sera is not known (Figure 4C). This implies that in the polyclonal dataset, a virus’s strong inhibition can be represented by a linear combination of other viral responses ($V_0 = a_1 \cdot V_1 + a_2 \cdot V_2 + \cdots$), and this same relationship holds in the monoclonal data so that $V_0$ can be predicted using the available measurements from $V_1$, $V_2$… The ability to bridge these two scales of the immune response implies that if a new viral variant emerges and is rapidly tested against a panel of sera, matrix completion can determine which previously characterized monoclonal antibodies potently inhibit the virus (Figure 4D).

Finally, we showed how combining datasets enables us to peer into the antibody responses at higher resolution by characterizing sera against an expanded set of viruses (Figure 6). In short, inter-table completion can rapidly sift through all missing antibody-virus combinations and prioritize efforts for groups seeking the strongest or weakest interactions.

A fertile and unexplored direction for future studies is to harness the low-rank patterns of virus inhibition to control an antibody response. In the Fonville human datasets, over 150 sera were described as a combination of 7–9 distinct phenotypes. By understanding how a specific phenotype can be elicited (e.g., potentially by vaccinating with a combination of viruses that are maximally inhibited in a low-rank pattern), sera could be guided to inhibit currently circulating strains. In essence, the low-rank antibody-virus patterns form a basis set of behaviors (akin to how a Fourier transform breaks a complex signal into a simple modes), through which the dynamics of the antibody response may be more readily understood.

Given the specificity and diversity of antibodies, it is worth speculating why antibody-virus interactions are well characterized in a low-dimensional space. First, we emphasize that matrix completion predicts the functional behavior of an antibody repertoire, which is far simpler than its composition (Snapkov et al., 2022). For example, sera from naive ferrets infected with the same virus often have similar HAI profiles (Kosikova et al., 2018; Arevalo et al., 2020), although the antibodies that each ferret makes are undoubtedly different. Second, even for more complex human responses, multiple sera have been found to only target a few epitopes (Lee et al., 2019a, 2019b), suggesting that these sera may be “approximately monoclonal.” Third,
Matrix completion can harness the low-rank behavior of antibody-virus interactions to greatly increase the amount of available data. In light of these results, it is worth considering how much data we actually have. Even when the viruses or antibodies characterized across multiple studies only partially overlap, patterns seen in one context can help inform other experiments. Indeed, from the vantage of matrix completion, differences in antibody and virus panels between experiments become an asset, enabling us to explore a greater fraction of the immune landscape and extrapolate the behavior of novel antibody-virus pairs.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **METHOD DETAILS**
  - Summary of available results and code
  - Nuclear norm minimization
  - Influenza: The Fonville 2014 Dataset
  - Influenza: The Vinh 2021 Dataset
  - Intra-Table Matrix Completion
  - Inter-Table Matrix Completion
  - Time and Memory Usage
  - Comparison with other Matrix Completion Algorithms
  - HIV-1: The Catnap Dataset
  - Number of Measurements Required for Accurate Matrix Completion

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2022.06.001.

**ACKNOWLEDGMENTS**

We would like to thank Jesse Bloom, John Huddleston, Heidi Klumpe, Richard Neher, and Armita Nourmohammad for their input on this manuscript. Tal Einav is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRQ 01-20). Brian Cleary was supported by a Merkin Institute Fellowship.

**AUTHOR CONTRIBUTIONS**

T.E. and B.C. conducted the research and wrote the paper.

**DECLARATION OF INTERESTS**

The authors declare no competing interests.
REFERENCES

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2021). Causal matrix completion. Preprint at arXiv. https://doi.org/10.48550/arxiv.2109.15145.

Agarwal, A., Verschueren, R., Diamond, S., and Boyd, S. (2018). A rewriting system for convex optimization problems. J. Control Decis. 5, 42–60. https://doi.org/10.1080/23307066.2017.1397554.

Amanu, I.J., Carlson, N.E., and Siftik, M.K. (2007). Duration of humoral immunity to common viral and vaccine antigens. N. Engl. J. Med. 357, 1903–1915. https://doi.org/10.1056/NEJMoa066092.

Arevalo, C.P., Le Sage, V., Bolton, M.J., Eilola, T., Jones, J.E., Kormuth, K.A., Nturibi, E., Balmaseda, A., Gordon, A., Lakdawala, S.S., et al. (2020). Original antigenic sin priming of influenza virus hemagglutinin stalk antibodies. Proc. Natl. Acad. Sci. USA 117, 17221–17227. https://doi.org/10.1073/pnas.1902321117.

Cai, Z., Zhang, T., and Wan, X.F. (2010). A computational framework for influenza antigenic cartography. PLoS Comput. Biol. 6, e1000949. https://doi.org/10.1371/journal.pcbi.1000949.

Candès, E.J., and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. J. ACM 58, 1–37. https://doi.org/10.1145/1970392.1970395.

Candès, E.J., and Recht, B. (2009). Exact matrix completion via convex optimization. Found. Comput. Math. 9, 717–772. https://doi.org/10.1007/s10208-009-9045-5.

Candes, E.J., and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inf. Theor. 56, 2053–2080. https://doi.org/10.1109/TIT.2010.2044061.

Cao, B., Adutwum, L.A., Oliynyk, A.O., Luber, E.J., Olsen, B.C., Mar, A., and Burik, J.M. (2018). How to optimize materials and devices via design of experiments and machine learning: demonstration using organic photovoltaics. ACS Nano 12, 7434–7444. https://doi.org/10.1021/acs.nanolett.8b04726.

Carter, D.M., Darby, C.A., Lefoley, B.C., Crevar, C.J., Arefin, T., Oomen, R., Anderson, S.F., Strugnell, T., Cortés-García, G., Vogel, T.U., et al. (2016). Design and characterization of a computationally optimized broadly reactive hemagglutinin vaccine for H1N1 influenza viruses. J. Virol. 90, 4720–4734. https://doi.org/10.1128/JVI.01315-15.

Cleary, B., and Regev, A. (2020). The power of convex relaxation: Near-optimal matrix completion. eLife 9, 561–573, July 20, 2022. https://doi.org/10.1016/j.cell.2021.08.048.

Georgiev, I.S., Doria-Rose, N.A., Zhou, T., Kwon, Y.D. Do, Stauep, R.P., Moquin, S., Chuang, G.Y., Louder, M.K., Schmidt, S.D., Attar-Trans, H.R., et al. (2013). Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. Science 340, 751–756. https://doi.org/10.1126/science.1233989.

Harvey, W.T., Benton, D.J., Gregory, V., Hail, J.P.J., Daniels, R.S., Bedford, T., Haydon, D.T., Hay, A.J., McCauley, J.W., and Reeve, R. (2016). Identification of low- and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A(H1N1) viruses. PLoS Pathog. 12, e1005526. https://doi.org/10.1371/journal.ppat.1005526.

Kong, R., Louder, M.K., Wagh, K., Bailier, R.T., DeCamp, A., Greene, K., Gao, H., Taft, J.D., Gazumyan, A., Liu, C., et al. (2015). Improving neutralization potency and breadth by combining broadly reactive HIV-1 antibodies targeting major neutralization epitopes. J. Virol. 89, 2659–2671. https://doi.org/10.1128/JVI.03136-14.

Koskova, M., Li, L., Radvák, P., Ye, Z., Wan, X.-F., and Xie, H. (2018). Imprinting of repeated influenza A/H3 exposures on antibody quantity and antibody quality: implications for seasonal vaccine strain selection and vaccine performance. Clin. Infect. Dis. 67, 1523–1532. https://doi.org/10.1093/cid/ciy027.

Kucharski, A.J., Lessler, J., Read, J.M., Zhu, H., Jiang, C.Q., Guan, Y., Cummings, D.A.T., and Riley, S. (2015). Estimating the life course of influenza A(H3N2) antibody responses from cross-sectional data. PLoS Biol. 13, e1002082. https://doi.org/10.1371/journal.pbio.1002082.

Lapedes, A., and Farber, R. (2001). The geometry of shape space: application to influenza. J. Theor. Biol. 212, 57–69. https://doi.org/10.1006/jtbi.2001.2347.

Lee, J., Paparotidis, P., Horton, A.P., Frühwirth, A., McDaniel, J.R., Jung, J., Gou, D.R., Hussein, D.A., Tanno, Y., Papass, L., et al. (2019a). Persistent antibody clonotypes dominate the serum response to influenza over multiple years and repeated vaccinations. Cell Host Microbe 25, 367–376.e5. https://doi.org/10.1016/J.CHEM.2019.01.010.

Lee, J.M., Eguia, R., Zost, S.J., Choudhary, S., Wilson, P.C., Bedford, T., Stevens-Ayers, T., Boechk, M., Hurl, A.C., Lakdawala, S.S., et al. (2019b). Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. eLife 8, e49324. https://doi.org/10.7554/eLife.49324.

Li, G.-M., Chiu, C., Wrammert, J., McCausland, M., Andrews, S.F., Zheng, N.-Y., Lee, J.-H., Huang, M., Xu, X., Edupuganti, S., et al. (2012). Pandemic H1N1 influenza virus induces a recall response in humans that favors broadly cross-reactive memory B cells. Proc. Natl. Acad. Sci. USA 109, 9047–9052. https://doi.org/10.1073/pnas.1119897109.

Liu, A., and Moitra, A. (2020). Tensor completion made practical. Adv. Neural Inf. Process. Syst. Preprint at arXiv. https://doi.org/10.48550/arxiv.2006.03134.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. 11, 2287–2322. https://pubmed.ncbi.nlm.nih.gov/21552465.

Mifflin, W. (2011). New methods for analyzing serological data with applications to influenza surveillance. Influenza Other Respir. Viruses 5, 206–212. https://doi.org/10.1111/j.1750-2698.2010.00192.x.

Perelson, A.S., Rong, L., and Hayden, F.G. (2012). Combination antiviral therapy for influenza: predictions from modeling of human infections. J. Infect. Dis. 205, 1642–1645. https://doi.org/10.1093/infdis/jis262.

Petrova, V.N., and Russell, C.A. (2018). The evolution of seasonal influenza viruses. Nat. Rev. Microbiol. 16, 47–60. https://doi.org/10.1038/nrmicro.2017.118.

Radhakrishnan, A., Stefanakis, G., Belkin, M., and Uhler, C. (2022). Simple, fast, and flexible framework for matrix completion with infinite width neural networks. Proc. Natl. Acad. Sci. USA 119, e2115064119. https://doi.org/10.1073/pnas.2115064119.

Recht, B. (2011). A simpler approach to matrix completion. J. Mach. Learn. Res. 12, 3413–3430. https://doi.org/10.5555/1953048.2185603.
Singer, A. (2008). A remark on global positioning from local distances. Proc. Natl. Acad. Sci. USA 105, 9507–9511. https://doi.org/10.1073/PNAS.0709842104.

Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmele, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2004). Mapping the antigenic and genetic evolution of influenza virus. Science 305, 371–376. https://doi.org/10.1126/science.1097211.

Snapkov, I., Chemigovskaya, M., Sinitcyn, P., Lê Quy, K., Nyman, T.A., and Greiff, V. (2022). Progress and challenges in mass spectrometry-based analysis of antibody repertoires. Trends Biotechnol. 40, 463–481. https://doi.org/10.1016/j.tibtech.2021.08.006.

Thompson, M.G., Naleway, A., Fry, A.M., Ball, S., Spencer, S.M., Reynolds, S., Bozeman, S., Levine, M., Katz, J.M., and Gaglani, M. (2016). Effects of Repeated Annual Inactivated Influenza Vaccination among Healthcare Personnel on Serum Hemagglutinin Inhibition Antibody Response to A/Perth/16/2009 (H3N2)-like Virus during 2010–11. Vaccine 34, 981–988. https://doi.org/10.1016/J.VACCINE.2015.10.119.

Vinh, D.N., Nhat, N.T.D., de Bruin, E., Vy, N.H.T., Thao, T.T.N., Phuong, H.T., Anh, P.H., Todd, S., Quan, T.M., Thanh, N.T.L. Le, et al. (2021). Age-seroprevalence curves for the multi-strain structure of influenza A virus. Nat. Commun. 12, 6680. https://doi.org/10.1038/s41467-021-26948-8.

Wagh, K., Seaman, M.S., Zingg, M., Fitzsimons, T., Barouch, D.H., Burton, D.R., Connors, M., Ho, D.D., Mascola, J.R., Nussenzweig, M.C., et al. (2018). Potential of conventional and bispecific broadly neutralizing antibodies for prevention of HIV-1 subtype A, C, and D infections. PLoS Pathog. 14, e1006860. https://doi.org/10.1371/journal.ppat.1006860.

Wrammert, J., Koutsonanos, D., Li, G.-M., Edupuganti, S., Sui, J., Morrissey, M., McCausland, M., Skountzou, I., Hornig, M., Lipkin, W.I., et al. (2011). Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J. Exp. Med. 208, 181–193. https://doi.org/10.1084/jem.20101352.

Yoon, H., Macke, J., West, A.P., Foley, B., Bjorkman, P.J., Korber, B., and Yusim, K. (2015). CATNAP: a tool to compile, analyze and tally neutralizing antibody panels. Nucleic Acids Res. 43, W213–W219. https://doi.org/10.1093/nar/gkv404.
STAR METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| Fonville influenza datasets | Fonville et al., 2014 | https://doi.org/10.1126/science.1256427 |
| Vinh influenza dataset | Vinh et al., 2021 | https://doi.org/10.1038/s41467-021-26948-8 |
| Software and algorithms | | |
| Matrix completion algorithm | This paper | https://doi.org/10.5281/zenodo.6623527 |

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tal Einav (tal.einav@gmail.com).

Materials availability
This study did not generate new materials.

Data and code availability
- **Source data statement:** This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table. The matrix-completed results for the Fonville, Vinh, and Catnap datasets are included in the supplemental information (Data S1).
- **Code statement:** All original code has been deposited in GitHub (Einav and Cleary, 2022) and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Summary of available results and code
We also include the code to perform matrix completion in both Mathematica and Python (https://github.com/cleary-lab/Inhibition-Neutralization-Completion). In Python, nuclear norm minimization takes fewer than 10 lines of code with the cvxpy package (Agrawal et al., 2018; Diamond and Boyd, 2016):

```python
matrix_incomplete = pandas.read_csv(filename)
mask = numpy.ones(matrix_incomplete.shape,dtype=numpy.int) # Mask for the missing data
mask[numpy.where(numpy.isnan(matrix_incomplete.values))] = 0
matrix_complete = cvxpy.Variable(shape=matrix_incomplete.shape)
objective = cvxpy.Minimize(mu * cvxpy.norm(matrix_complete, "nuc") + cvxpy.sum_squares(cvxpy.multiply(mask, matrix_complete - matrix_incomplete)))
problem = cvxpy.Problem(objective, )
problem.solve(solver=cvxpy.SCS)
```

Nuclear norm minimization
One intuitive way to fill in missing values within a matrix is to minimize matrix rank, which describes all serum and virus behavior using as few patterns as possible (with the smallest basis set). Figure S1 shows two simple examples of a rank 1 and rank 2 matrix, each with two missing values. For these simple and noise-free examples, we can identify the underlying basis set and fill in the missing values. However, since minimizing matrix rank is NP-hard, more complex and noisy matrices require a more computationally efficient method.

To illustrate nuclear norm minimization on each of these simple matrices $M$, we compute their singular value decomposition, $M=USV^\top$, and minimize the sum of singular values $\text{Tr}(\Sigma)$. As shown in Figure S1, the resulting solution is a good proxy for minimizing matrix rank. When larger matrices have many missing values, this minimization generalizes to a high-dimensional search through parameter space, and in this regime minimizing the nuclear norm is far more efficient than minimizing matrix rank.
Influenza: The Fonville 2014 Dataset
We gathered antibody-virus HAI measurements from the supplementary information of Fonville 2014 that contains the following six studies:

- Study 1 [Fonville Table S1]: Ferret study; 35 sera measured against 74 viruses
- Study 2 [Fonville Table S3]: Human infection study from 2007-2012; 324 sera measured against 57 viruses
- Study 3 [Fonville Table S5]: Human vaccination study from 1997; 212 sera measured against 70 viruses
- Study 4 [Fonville Table S6]: Human vaccination study from 1998; 256 sera measured against 70 viruses
- Study 5 [Fonville Table S13]: Human vaccination study from 2009; 160 sera measured against 20 viruses
- Study 6 [Fonville Table S14]: Human vaccination study from 2010; 160 sera measured against 20 viruses

Throughout this work, we withheld subsets of the available measurements (grayscale in Figure S2) and compared our predictions with these measurements. For intra-table completion, we did not include in the matrix completion statistics any columns or rows that were entirely removed, since we cannot predict the behavior of an entirely unmeasured virus or serum using a single dataset.

Each serum sample only appeared in a single study, whereas all viruses were included in at least two studies. Cumulatively, the six studies contained 1147 distinct serum samples and 81 distinct viruses, with the distribution of viruses across the six studies shown in Figure S5. The raw data annotated with the virus names is shown in Figure S2 and included in the GitHub repository associated with this paper.

Influenza: The Vinh 2021 Dataset
The Vinh dataset measures 24,000 sera against the following six H3N2 viruses:

- Virus 1 = A/Victoria/361/2011
- Virus 2 = A/Victoria/210/2009
- Virus 3 = A/Brisbane/10/2007
- Virus 4 = A/Wisconsin/67/2005
- Virus 5 = A/Wyoming/3/2003
- Virus 6 = A/Hanoi/EL201/2009, short name=HN201/2009, 7 amino acids away from A/Victoria/210/2009 and virus #6 (full name=A/Bilthoven/16190/1968, short name=BI/16190/68, 3 amino acids different from A/Victoria/210/2009). For our analysis, we associate the data across all studies from these two nearly equivalent strains. The Vinh panel also includes five H1N1 viruses, but we ignore those strains because the Fonville virus panel only contains H3N2 viruses.

Intra-Table Matrix Completion
Consider a table of measurements $M$ where entry $M_{jk}$ represents the interaction between antibody (or serum) $j$ and virus $k$. $M_{jk}$ either represents the hemagglutination inhibition (HAI) titer from the Fonville influenza study or the inverse of the 50% inhibitory concentration $(IC_{50})$ for monoclonal HIV-1 antibodies from Catnap. Thus, potent antibody/serum interactions are always denoted by larger values.

Influenza HAI is measured using progressive 2-fold dilutions, so that titers can equal “<10”, 10, 20, ..., 1280, or “≥1280.” We first replaced lower or upper bounds (“<10” → 5 and “≥1280” → 2560) and then took $\log_{10}$ of all titers. To prevent the error from large values from artificially dominating the completion algorithm, we subtracted the minimum value $m = \min_m(M_{jk})$ from all entries in $M$, performed matrix completion, and then added $m$ back to obtain the complete matrix with log titers (to transform to the original un-logged titers, the matrix completed value can be exponentiated by 10). We note that as an alternative, the mean of $M$ could be subtracted to center the data in log-space; in this work, we chose to subtract the minimum titer in the majority of cases because it results in slightly better predictions of the larger titer values (which are often the most important measurements). However, we used mean-subtraction for the Fonville-Vinh inter-table completions because the two datasets had markedly different distributions of titers, and mean-subtraction led to more accurate predictions.

The goal of matrix completion is to find a filled-in low-rank matrix $\tilde{M}$ such that $M_{jk} = \tilde{M}_{jk}$ for all pairs $(j,k)$ that were measured. Note that these two criteria are at odds with one another, since maintaining exact fidelity with the data may result in a matrix $\tilde{M}$ with larger rank. We control the relative importance of these two goals with a single hyperparameter $\lambda$ described below.

Although minimizing matrix rank is NP-hard, minimizing the convex envelope of the rank (the nuclear norm $\|M\|_*$) is computationally efficient. Hence, the matrix $\tilde{M} \in R^{n \times m}$ is completed as

$$\min \{ \tilde{M} \in R^{n \times m} : \|M\|_* + \lambda \|M_{\text{measured}} - M_{\text{measured}}\|_2^2 \}$$

where $(M_{\text{measured}})_{jk} = M_{jk}$ whenever entry $(j,k)$ was measured and 0 otherwise, with an analogous definition for $M_{\text{measured}}$. The first term, $\|M\|_*$, minimizes the nuclear norm (and hence the rank) of $M$, whereas the second term $\|M_{\text{measured}} - M_{\text{measured}}\|_2^2$ is a quadratic penalty for deviating from the measured values (the prefactor $\lambda$ is discussed below). Such minimizations have a known tendency to
shrink values towards zero, so that the best-fit line of the measured vs predicted titers has a slightly larger slope than the perfect experiment-theory match (purple vs black lines in the scatterplots of Figures 2B and 3B). For the influenza data where \( \log_{10}(\text{titers}) \) are matrix completed, the predicted values tend to be smaller than the measured titers; for the HIV-1 antibody data where the \(-\log_{10}(\text{IC}_{50})\) values are matrix completed, the predicted values tend to be greater than the measured titers because of the negative (see the HIV-1: The Catnap Dataset methods section below).

The hyperparameter \( \lambda \) dictates the relative importance of the two terms in the minimization. Large values of \( \lambda \) indicate strong trust in the measured values and would result in higher-rank completions that remain highly faithful to the data. Smaller values of \( \lambda \) would be appropriate when the data is noisy and will result in a lower-rank matrix that can deviate from the measurements. In this work, we used \( \lambda=1 \) in our analyses. With this \( \lambda=1 \) value, over 95% of the completed Fonville titers were within 10% of the measured values, keeping high fidelity with the experimental values.

When we varied this hyperparameter, we found a very weak dependence on \( \lambda \); for example, choosing \( \lambda=1, \lambda=10, \text{or} \lambda=100 \) led to similar intra-table completion curves, although lower values of \( \lambda=0.1 \) or \( \lambda=0.01 \) resulted in an “overly-smoothed” matrix whose rank was too low (Figure S13). Thus, any choice of \( \lambda \geq 1 \) yields comparable matrix completion accuracy (and the Python code in the GitHub repository uses \( \lambda=1 \) by default). In the limit \( \lambda \to \infty \), the complete matrix must exactly match the input matrix at all measured values, and the missing values are filled in to minimize the matrix rank without requiring any hyperparameters (this is implemented in the Mathematica code in the GitHub repository). Lastly, we note that as more datasets are added, there may be a more complex dependence between \( \lambda \) and the accuracy of matrix completion, and in such cases the optimal value of \( \lambda \) can be determined through cross-validation.

Because the predicted and measured titers span multiple decades and are presented on log-log plots, both Pearson’s \( r^2 \) and RMSE were computed on the \( \log_{10} \) data. The RMSE can be recast into un-logged titers by exponentiating by 10 (e.g., an RMSE of 0.2 in \( \log_{10} \) space corresponds to \( 10^{0.2} \approx 1.6 \)-fold error). The dimensionality of the data was computed on the titers mean-centered in log space, \( M_{\text{centered}} = \log_{10}(M) - \text{mean}(\log_{10}(M)) \) to prevent a change in the units (e.g., measuring an IC\textsubscript{50} in \( \mu \text{g/mL} \) versus Molar) from affecting the dimensionality. The dimensionality \( r \) of \( M_{\text{centered}} \) is defined as the minimum number of singular values of \( M_{\text{centered}} \) that explain 95% of its variance (the minimum \( r \) such that 0.95 \( \leq \sum_{i \leq \lambda} \frac{\sigma_i^2}{\sigma_1^2} \)).

Inter-Table Matrix Completion

Computationally, inter-table completion works identically to intra-table completion, except that virus and serum panels must first be combined. Note that some of the Fonville viruses had different capitalization or punctuation in different tables (e.g., “HN196/2009” and “JO/33/94” in Fonville Table S1 and their counterparts “HN196/2009” and “A/JOHANNESBURG/33/94” in Fonville Table S3) that we unified in the supplementary data files. Any virus that was not present in a study was defined as missing for all sera in that study (e.g., “A/AUCKLAND/20/2003” was contained in Studies 1, 3, and 4, so when combining all Fonville data, its titers were defined as “Missing” against sera in Studies 2, 5, and 6).

In the few cases when all predicted or measured \( \log_{10}(\text{titers}) \) withheld from a dataset have nearly identical values, the correlation coefficient \( r^2 \) is artificially deflated. For example, if five (predicted, measured) titers for a virus were \( \log_{10} \) of \((10,10), (10,10), (10,10), (10,20), (20,10))\), then any linear dependence is masked by the small range of the predictions and measurements. This leads to a small Pearson’s \( r^2=0.06 \), even though the predictions were perfect for the first three points and only 2-fold off for the final two points. In contrast, when measurements span a wide dynamic range with this same error distribution, such as \( \log_{10} \) of \((10,10), (20,20), (40,40), (80,160), (160,80))\), their resulting Pearson’s \( r^2=0.81 \), matching the typical intuition for this metric. Hence, when all predicted or measured \( \log_{10}(\text{titers}) \) withheld had a standard deviation \( \leq 0.2 \), we excluded such cases from the Pearson’s \( r^2 \) analysis in Figures 4B, S5, and S7. We note that such cases only occurred with inter-table completion for viruses that showed minimal HAI against all sera (Figure 4B) and never occurred for intra-table completion (Figures 2 or 3).

When matrix completing between the Fonville and Vinh datasets (Figure 6), we first intra-table completed the Fonville data, then trimmed it to either the Vinh virus panel (6 viruses) or the extended virus panel (6+17 viruses), and then inter-table completed the Vinh measurements. This first intra-table completion step is important, because many Fonville sera were not measured against all six Vinh viruses, and hence these sera would have been trimmed away without contributing to the Vinh completions. Performing this first matrix completion harnesses the full array of measurements in the Fonville dataset to make maximally informed predictions about the Vinh viruses. Moreover, when combining measurements from different studies (e.g., from Fonville and Vinh studies), the accuracy of matrix completion improved when we separately mean-centered both datasets in log-space (i.e., divide all titers by the mean of the measured values; in contrast, we min-centered all values during intra-table completion). Mean-centering can adjust for different protocols across labs or for systematic shifts in the data. In addition, prediction accuracy improved when only a single additional viral virus was imputed (i.e., predicting the single withheld Vinh virus without the 17 additional Fonville strains). Thus, when we predicted the additional Fonville strains, we predicted their titers one at a time.

Time and Memory Usage

The nuclear norm described in this work scales polynomially with the size of the matrix. While the time required will vary with the dimensions of the matrix, a handy rule of thumb is that a matrix with 100 viruses and 100 sera takes \( \sim 1 \) second to complete.
Below are some benchmarks for the time and memory usage of matrix completion running Python on a high-performance computing machine:

- (2000 sera) x (6 viruses): 0.5 hours, 5 GB memory
- (2000 sera) x (80 viruses): 0.5 hours, 6 GB memory
- (6000 sera) x (6 viruses): 8 hours, 30 GB memory
- (6000 sera) x (80 viruses): 11 hours, 32 GB memory

The supplementary Mathematica notebook uses a Soft-Impute method (Mazumder et al., 2010) similar to a variant of robust PCA (Candès et al., 2011). This approach yields nearly identical results to nuclear norm minimization, but is over 100x faster. Below are these same benchmarks for the time and memory usage running Mathematica on a standard laptop:

- (2000 sera) x (6 viruses): 3 seconds, 7 GB memory
- (2000 sera) x (80 viruses): 30 seconds, 70 GB memory
- (6000 sera) x (6 viruses): 10 seconds, 20 GB memory
- (6000 sera) x (80 viruses): 50 seconds, 240 GB memory

**Comparison with other Matrix Completion Algorithms**

As a point of comparison, we replicated our antibody-virus analysis using three other imputation approaches: mean imputation, $k$ nearest virus neighbors, and $k$ nearest serum neighbors. For the latter two algorithms, we used the hyperparameters $k=15$ or $k=20$, respectively, that minimized the RMSE during intra-table completion. We found that nuclear norm minimization outperformed all three algorithms (Figure S12).

Moreover, we tested a recently-published method of causal matrix completion that was developed to specifically deal with scenarios in which data are not missing at random, and in which the availability of data may be correlated with the outcome of the experiment (Agarwal et al., 2021). This could be relevant, for instance, in matrix completion of the HIV-1 Catnap data through a given date, since measurements made by the community were not randomly chosen, but rather chosen to be informative for a specific study. However, we found that the causal matrix completion approach performed no better than nuclear norm minimization in this setting ($r^2=0.43$ using causal matrix completion versus $r^2=0.45$ with nuclear norm minimization, when predicting observations made in 2019 or later based on available data through 2018). In the smaller Fonville analyses, we found that causal matrix completion performed substantially worse than nuclear norm minimization.

**HIV-1: The Catnap Dataset**

The complete list of Catnap antibody-virus measurements is updated each month on the Downloads page (https://www.hiv.lanl.gov/components/sequence/HIV/neutralization/download_db.comp; click the file labeled “Assay”). We split each antibody-virus pair into one of four categories based on the antibody name and the IC$_{50}$ and ID$_{50}$ values:

1. Polyclonal sera ($n=40$) did not have IC$_{50}$ values but instead had ID$_{50}$ values (the dilution at which a virus was 50% neutralized)
2. Antibody mixtures ($n=44$) had a “+” in their antibody names to distinguish the monoclonal antibodies in each mixture
3. Multispecific antibodies ($n=50$) had an “/” in their antibody name to denote the different monoclonal antibodies from which they were composed
4. Monoclonal antibodies ($n=373$) contained all remaining entries

For simplicity, our Catnap analysis was restricted to the monoclonal antibodies (category #4 above) and polyclonal sera (category #1). We replaced values bounded from above by half the value of the bound and values bounded from below by double their value (e.g., “<20” → 10, “>40” → 80), and when multiple antibody-virus measurements were available we took their geometric mean. We note that these assumptions, especially our treatment of the bounded values, may influence the matrix completion results given that the IC$_{50}$ values in Catnap range from 10$^{-5}$ to 344 µg/mL, while the upper bounds varied between 10$^{-3}$ and 10$^{0}$ µg/mL.

The only minor difference between the Catnap HIV-1 analysis and the Fonville influenza analysis accounted for the fact that potent antibodies have low IC$_{50}$ values in Catnap whereas potent sera have a high HAI titer in Fonville. Therefore, we completed the matrix of -log$_{10}$(IC$_{50}$ values) from Catnap monoclonal antibody data, whereas we completed the matrix log$_{10}$(HAI titers) for Fonville and log$_{10}$(ID$_{50}$) for Catnap polyclonal serum data. Note that in all cases, we represented the weak antibody/sera by smaller values of log$_{10}$IC$_{50}$ values in Catnap range from 10$^{-5}$ to 344 µg/mL, while the upper bounds varied between 10$^{-3}$ and 10$^{0}$ µg/mL.

**Number of Measurements Required for Accurate Matrix Completion**

When missing values are randomly sampled across a matrix for intra-table completion, the theoretical bound in Equation 1 quantifies the number of measurements ($m$) required to accurately reconstruct a matrix based on the number of antibodies or viruses $n$ (equal to max(# of antibodies, # of viruses)), the matrix rank $r$ (here, determined as the rank required to explain 95% of the variance of the completed matrix), the coherence $\mu$ (calculated from the complete matrix as explained below), and a constant $C$ (inferred from...
the Fonville RMSE curves where $m$ can be determined empirically). Together, these values quantify how difficult it is to complete a given dataset.

To compute $\mu$ for any study (Candès and Recht, 2009; Ndifon, 2011), we take all rows/columns of the data matrix with $\geq 50\%$ of entries measured and calculate the completed matrix $M$ whose dimensions we denote by $n_1 \times n_2$. (Note that using the rows/columns with few measurements leads to unstable results for highly incomplete datasets such as Catnap and are hence dropped.) We compute the rank $r$ of $M$ as the minimum number of singular values of $M$ that explains $95\%$ of its variance. Using the singular value decomposition $M=U\Sigma V^T$, we form the orthogonal projection onto $M$ using the first $r$ eigenvectors of $U$ or $V$. For example, $P_U=\sum_{1 \leq j \leq r} u_j u_j^T$ where $u_j$ is the $j^{th}$ column of $U$, with an analogous expression for $V$. The coherence of $M$ is given by

$$\mu = \max \left( \frac{n_2}{r} \max_{1 \leq j \leq n} \| P_U e_j \|_2^2, \frac{n_2}{r} \max_{1 \leq k \leq n} \| P_V e_k \|_2^2 \right).$$

(Equation 2)

where $e_j$ and $e_k$ represent standard basis elements. The resulting values are given in Table S2.

To compute the value of the constant $C$ in the Fonville studies, we first quantify the minimum number of measurements $m$ at which the slope of the RMSE curves in Figure 2 becomes greater than $-0.5$ (indicating the sufficient number of measurements for that study). We then solve for $C$ in Equation 1, obtaining a value for each study (Table S2). Below, we use the averaged value, $C=0.18$.

To predict the number of measurements needed to bottom-out the Catnap RMSE and $r^2$ curves in Figures 2D and 3D, we utilize the theoretical bound in Equation 1 that a matrix can be accurately completed when the number of (randomly observed) measurements is greater than $C m r n \log(n)$, where $n=933$ is the larger dimension of the Catnap matrix and $r=23$ equals the rank of the completed matrix (Candès and Recht, 2009). To validate this result, we apply this same approach to the Catnap polyclonal sera dataset and predict that the RMSE curve will bottom-out when $41\%$ of measurements are observed. Since the polyclonal serum dataset has approximately half of all serum-virus interactions measured, we can directly test this prediction by performing intra-table completion using different fractions of observed data — in doing so, we find that the RMSE curves bottoms-out when $44\%$ of measurements are observed, close to our predicted value.