Massively Increasing the number of Antibody-Virus Interactions across Studies

Authors: Tal Einav[1]*, Rong Ma[2]*

Affiliations:
[1]Basic Sciences Division and Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

[2]Department of Statistics, Stanford University, Stanford, California, United States of America

*Authors contributed equally to this work

Correspondence should be addressed to teinav@fredhutch.org

Abstract
A central challenge in every field of biology is to use existing measurements to predict the outcomes of future experiments. In this work, we consider the wealth of antibody inhibition data against variants of the influenza virus. Due to this virus's genetic diversity and evolvability, the variants examined in one study will often have little-to-no overlap with other studies, making it difficult to discern common patterns or unify datasets for further analysis. To that end, we develop a computational framework that predicts how an antibody or serum would inhibit any variant from *any other study*. We use this framework to greatly expand 7 influenza datasets utilizing hemagglutination inhibition, validating our method upon 200,000 existing measurements and predicting more than 2,000,000 new values along with their prediction uncertainties. This data-driven approach does not require any information beyond each virus's name and measurements, and even datasets with as few as 5 viruses can be expanded, making this approach widely applicable. Future influenza studies using hemagglutination inhibition can directly utilize our curated datasets to predict newly measured antibody responses against ≈80 H3N2 influenza viruses from 1968-2011, whereas immunological studies utilizing other viruses or a different assay only need to find a single partially-overlapping dataset to extend their work. In essence, this approach enables a shift in perspective when analyzing data from "what you see is what you get" into "what anyone sees is what everyone gets."

Introduction

Our understanding of how antibody-mediated immunity drives viral evolution and escape relies upon painstaking measurements of antibody binding, inhibition, or neutralization against variants of concern (Petrova and Russell, 2017). For a rapidly evolving virus such as influenza, the specific variants examined in one study will often have little-to-no overlap with other studies, making it difficult to directly compare antibody responses or augment results using existing data. This lack of cross-talk hampers our ability to comprehensively characterize viral antigenicity, predict the outcomes of viral evolution, and make critical decisions such as whether to update the annual influenza vaccine (Morris et al., 2018).

In this work, we develop a new cross-study matrix completion algorithm that leverages patterns in antibody-virus inhibition data to accurately infer unmeasured interactions. Specifically, we demonstrate that multiple datasets can be combined to predict the behavior of viruses that were entirely absent from one or more datasets (*e.g.*, Figure 1A, predicting values for the green viruses in Dataset 2 and the gray viruses in Dataset 1). Whereas past efforts could only predict values for partially-observed viruses within a single dataset (*i.e.*, predicting the red squares for the green/blue viruses in Dataset 1 or the blue/gray viruses in Dataset 2) (Cai et al., 2010; Ndifon, 2011; Einav and Cleary, 2022), here we predict data for viruses that do not yet have a single measurement in a dataset.

From this vantage, even when each dataset is individually complete, there are still unmeasured interactions that can be inferred by combining studies. For example, each dataset we examine in this work measured 60–100% of interactions between its specific panel of viruses and antibody responses (sera). Yet when we combine datasets and ask how each serum inhibits any virus from any study, fewer than 10% of interactions are measured. The structure of these missing values violates the underlying assumption of most existing matrix completion methods, which assume that missing entries are randomly distributed (Candes and Recht, 2009; Cai et al., 2010; Candes and Tao, 2010; Candes and Plan, 2010; Keshavan et al., 2010). In contrast, we construct our framework to harness this structure, enabling us to predict over 2,000,000 new values comprising the remaining 90% of interactions.

Algorithms that predict the behavior of large virus panels are crucial because they render the immunological landscape in higher resolution, helping to reveal which viruses are potently inhibited and which escape antibody immunity (Morris et al., 2018). Even polyclonal human sera that strongly neutralizes one virus may exhibit 10x weaker neutralization against a virus with one additional mutation (Lee et al., 2019b). Given the immense diversity and rapid evolution of viruses, it behooves us to pool together measurements and build a more comprehensive description of serum behavior.

The key missing feature we develop that makes matrix completion practical across studies is error quantification. Although matrix completion has been extensively studied since its inception 20 years ago, there is a scarcity of methods that quantify the uncertainty of predicted values for general patterns of missing data (Chen et al., 2019). In this study, we examine the highly-structured scenario where entire columns (representing viruses) are missing, and we quantify the *individual error* of each prediction based on the specific datasets used to infer the missing values. As a result, users can focus

on high-confidence inferences (*e.g.*, those with ≤4-fold error) or search for additional datasets that would further reduce this uncertainty.

To demonstrate that our error quantification is robust, we examine how different combinations of input datasets result in more accurate predictions. We show that prediction uncertainty only decreases whenever more input datasets are used in our analysis, quantifying the advantage of combining multiple datasets.

In addition, we introduce the notion of *transferability* between datasets, that is, how accurately virus behavior in Dataset *X* predicts the values in Dataset *Y*. This enables us to quantify how measurements in ferrets transfer into humans, how infection studies relate to vaccination studies, and how time affects transferability. As future studies probe increasingly-diverse antibody responses, transferability maps can quantify and visualize their relationship to previous datasets.

Our results push beyond traditional applications of matrix completion in several key ways: (*i*) By quantifying error, we open up a practical implementation to combine datasets, greatly increasing the scope and utility of completion algorithms. (*ii*) Existing antibody-virus datasets can be unified so that each serum is predicted against the same extended virus panel. This massive expansion of existing data not only provides fine-grained resolution into these antibody responses but also enables an unprecedented direct comparison between studies utilizing distinct virus panels. (*iii*) Ongoing studies can leverage matrix completion to save time and resources by measuring a substantially smaller subset of viruses instead of the full panel. In particular, our approach can determine which subset of viruses will be maximally informative.

Although this work focuses on antibody-virus inhibition measurements, it readily generalizes to other viruses, other assays (*e.g.*, binding or neutralization), and even other applications involving intrinsically low-dimensional datasets.

Results

The Low Dimensionality of Antibody-Virus Interactions Empowers Matrix Completion

Given the vast diversity of antibodies, it is easy to imagine that each person's serum is unique, and that one serum response cannot inform the behavior of another. Indeed, factors such as age, geographic location, frequency/type of vaccinations, and infection history shape the antibody repertoire and influence how it responds to a vaccine or to a new viral threat (Kim et al., 2012; Fonville et al., 2014; Thompson et al., 2016; Gouma et al., 2020; Fox et al., 2022).

Yet much of the heterogeneity of antibody responses collapses when we consider their functional behavior such as binding, inhibition, or neutralization against viruses. Previous work has shown that antibody-virus inhibition data are intrinsically low-dimensional (Lapedes and Farber, 2001), which spurred applications ranging from antigenic maps to the recovery of missing values from partially observed data (Smith et al., 2004; Cai et al., 2010; Ndifon, 2011; Einav and Cleary, 2022). Yet these efforts have almost exclusively focused on individual datasets, circumventing the difficulties in predicting values across studies.

To appreciate the magnitude of this challenge, note that it is only possible to predict measurements for a virus in Dataset 1 (*e.g.*, the virus-of-interest in Figure 1A, boxed in gold) using Dataset 2 provided that: (1) The virus-of-interest is related to at least some of the viruses present in both studies (the other blue viruses), (2) the transferability of information between Dataset 1 and Dataset 2 is quantified, accounting for potential systematic differences in the data, and (3) there are a sufficient number of antibody responses and viruses in Dataset 2 to infer the missing values. If these conditions are met, it intuitively makes sense that the patterns between the blue viruses inferred from Dataset 2 can predict the virus-of-interest in Dataset 1.

In this work, we first demonstrate the accuracy of matrix completion by withholding *all* measurements from one virus in one dataset [Figure 1A, gold boxes] and using the other datasets to generate *predictions±errors*, where each *error* quantifies the uncertainty of a *prediction*. We emphasize that this error estimation is vital, since in some scenarios it may be impossible to accurately predict an antibody response (*e.g.*, measurements of viruses from 2000-2010 may not be able to predict a distant virus from 1970). Hence, both accurate predictions with small uncertainty as well as inaccurate predictions with large uncertainty demonstrate that our approach is faithful and robust.

In the following sections, we develop a matrix completion algorithm that incorporates the features described above. We then analyze 7 large serological studies containing hemagglutination inhibition (HAI) measurements for human or animal antibody responses. We first validate our approach by withholding and predicting known data and then apply matrix completion across these studies to greatly extend their measurements.

Cross-Study Matrix Completion using a Random Forest Algorithm

We first examine how virus behavior can be predicted between two studies before considering multiple studies. Figure 1 and Algorithm 1 summarize what we call *leave-one-out* analysis, where a virus-of-interest V_0 (blue virus boxed in gold) is completely withheld and its values are predicted from Dataset 2→1.

To begin, we enumerate all viruses measured in both studies as V_1, $V_2 \ldots V_n$ (n=7 other blue viruses in Figure 1) and quantify the relationship between V_1-V_n and V_0. To that end, we form a training set by sampling with replacement 5 viruses (features) from V_1-V_n and 30% of the antibody responses (samples) in Dataset 2 [Figure 1A, purple boxes]. We then train a decision tree model to predict V_0 using this feature and sample set, and we quantify its root-mean-squared error (RMSE) σ_{Training} through cross-validation on the 70% of unsampled antibody responses [Figure 1B].

Since we do not know which of the viruses V_1-V_n will best characterize V_0, we train multiple decision trees using different subsets of features and samples. Using the 5 trees with the lowest σ_{Training}, we predict the values of V_0 for each serum in Dataset 1 by averaging the value μ and error σ_{Training} from all 5 trees [Figure 1B]. This approach of averaging over multiple trees, called a random forest algorithm, avoids the tendency of individual decision trees to overfit the training set.

One potential pitfall of this approach is that the estimated error σ_{Training} derived from Dataset 2 will almost always underestimate the true RMSE for these predictions (σ_{Actual}) in Dataset 1, since the

antibody responses in both studies may be very distinct (*e.g.*, responses measured in different years or collected from people/animals with different infection histories). Hence, the relationships between viruses in Dataset 2 may not reflect the relationships between these same viruses in Dataset 1.

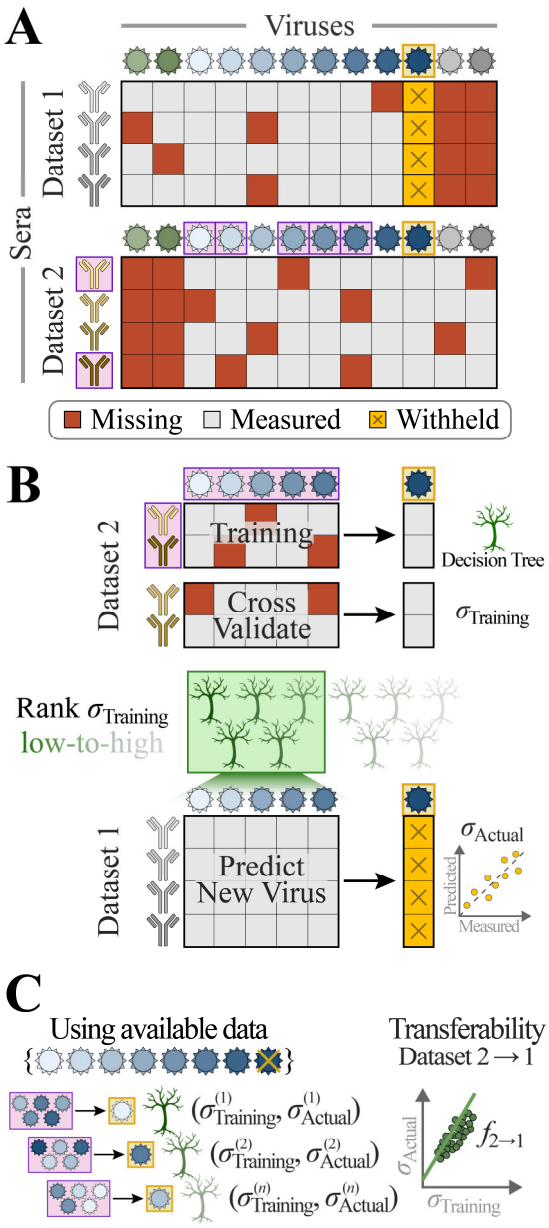

Figure 1. Combining datasets to predict the values and uncertainty for entirely missing viruses. (A) Schematic of two studies measuring different antibody responses against some overlapping viruses (shades of blue) and some unique viruses (green/gray). Studies may have different fractions of missing values (dark-red boxes) and measured values (gray). To demonstrate that virus behavior can be inferred across studies, we withhold all measurements from one overlapping virus in Dataset 1 (gold squares) and predict those values. We repeat this process, withholding each virus in every dataset. (B) Virus behavior in Dataset 1 is predicted using patterns from Dataset 2. To train our model, we choose a random subset of antibody responses and overlapping viruses (both boxed in purple) and infer the relationship between this virus subset and the withheld virus in Dataset 2 using a decision tree model. This decision tree is cross-validated against the unused samples in Dataset 2 by computing its root-mean-squared error (RMSE, denoted by σ_{Training}). Predictions for the withheld virus in Dataset 1 are given by the average from the 5 trees with the lowest error, and the RMSE of these predictions is denoted by σ_{Actual}. (C) To estimate the

prediction error σ_{Actual} in Dataset 1 (which we cannot directly compute because these values are withheld), we create multiple additional decision trees between Dataset 2→1 using all available measurements (but excluding the withheld virus). For these decision trees, we can access the cross-validation error $\sigma_{Training}$ in Dataset 2 as well as the true error σ_{Actual} in Dataset 1, thereby quantifying the transferability relationship $f_{2\rightarrow1}$. We apply this relation to the decision trees in Panel B to estimate the prediction error for the withheld virus in Dataset 1.

To correct for this effect, we quantify the transferability map $f_{2\rightarrow1}(x)$ between Dataset 2 and Dataset 1, so that for the majority of decision trees, $f_{2\rightarrow1}(\sigma_{Training}$ from Dataset 2$)\geq\sigma_{Actual}$ in Dataset 1. To that end, we create additional decision trees, this time restricting our analysis to viruses V_1-V_n and predicting each in turn [Figure 1C, Algorithm 2]. Since none of these viruses were withheld, we can directly compare their estimator $\sigma_{Training}$ and true σ_{Actual} to infer $f_{2\rightarrow1}$, which we found was well-characterized by a simple linear relationship [Figure S1, Methods]. Finally, we use this relation to compute the prediction error for V_0 in Dataset 1, $\sigma_{Predict}=f_{2\rightarrow1}(\sigma_{Training})$, for the top 5 decision trees. In this way, both the values and errors are inferred using a generic, data-driven approach that can be applied to diverse datasets.

Leave-One-Out: Inferring the Behavior of Viruses without a Single Measurement

To assess matrix completion across studies, we applied it in three increasingly difficult scenarios: (1) between two highly-similar human vaccination studies, (2) between a human infection and human vaccination study, and (3) between a ferret infection and human vaccination study. We expect that prediction accuracy will decrease as the datasets become more distinct, resulting in a larger error (σ_{Actual}) and larger estimated uncertainty ($\sigma_{Predict}$).

For these predictions, we utilized the Fonville influenza dataset consisting of six studies: four human vaccination studies ($Dataset_{Vac,1\text{-}4}$), one human infection study ($Dataset_{Infect,1}$), and one ferret infection study ($Dataset_{Ferret}$) (Fonville et al., 2014). In each study, sera were measured against a panel of H3N2 viruses using hemagglutination inhibition. Collectively, these studies contained 81 viruses, and each virus was measured in at least two studies.

We first predicted values for the virus V_0=A/Auckland/5/1996 in the most recent vaccination study ($Dataset_{Vac,4}$) using data from another vaccination study ($Dataset_{Vac,3}$) carried out in the preceding year and in the same geographic location [Table 1]. After training our decision trees, we found that the two studies had near-perfect transferability ($\sigma_{Predict}=f_{Vac,3\rightarrow Vac,4}(\sigma_{Training})\approx\sigma_{Training}$), suggesting that there is no penalty in extrapolating virus behavior between these datasets. More precisely, if there exist five viruses V_1-V_5 that can accurately predict V_0's measurements in $Dataset_{Vac,3}$, then V_1-V_5 will predict V_0 equally well in $Dataset_{Vac,4}$.

Indeed, we found multiple such decision trees, with which we predicted V_0 together with the uncertainty of $\sigma_{Predict}$=1.8-fold, meaning that a measurement with a predicted value x is expected to lie between x/1.8 and x·1.8 [top panel in Figure 2A, gray bands represent $\sigma_{Predict}$]. Moreover, this estimated uncertainty closely matched the true RMSE σ_{Actual}=1.6-fold. To put these results into perspective, the HAI assay has roughly 2-fold error (*i.e.*, repeated measurements differ by 2-fold 50% of the time and by 4-fold 10% of the time, Methods), implying that these predictions are as good as possible given experimental error.

When we inferred every other virus between Datasets$_{Vac,3\rightarrow Vac,4}$, we consistently found the same highly accurate predictions $\sigma_{Predict}\approx\sigma_{Actual}\approx$2-fold [Figure S2A]. As an alternative way of quantifying error, we plotted the distribution of predictions within 0.5, 1.0, 1.5… standard deviations from the measurement, which we compare against a folded Gaussian distribution [Figure 2A, bottom]. For example, 81% of predictions were within one standard deviation, somewhat larger than the 68% expected for a Gaussian, which suggests that the prediction error was slightly overestimated.

We next predicted values for V_0=A/Netherlands/620/1989 between a human infection and vaccination study (Dataset$_{Infect,1\rightarrow Vac,4}$). In this case, the predicted values were also highly accurate with true error σ_{Actual}=2.2-fold [Figure 2B; remaining viruses predicted in Figure S2B]. When quantifying the uncertainty of these predictions, we found far less transferability of virus behavior ($\sigma_{Predict}=f_{Infect,1\rightarrow Vac,4}(\sigma_{Training})\approx 2.8\sigma_{Training}$, where a larger slope indicates less transferability, see Methods), and hence we overestimated the prediction error as $\sigma_{Predict}$=4.0-fold. Lastly, when we predicted values for V_0=A/Perth/27/2007 between a ferret infection and human vaccination study (Dataset$_{Ferret\rightarrow Vac,4}$), our predictions had increased true error σ_{Actual}=4.6-fold [Figure 2C], and poor transferability again led to a larger estimated prediction error $\sigma_{Predict}$=6.6-fold. Importantly, this overestimation is purposefully built into the transfer functions $f_{X\rightarrow Y}$ whenever Datasets X and Y exhibit very noisy or highly disparate behaviors, so that we err on the side of overestimating rather than underestimating uncertainty whenever prediction is intrinsically difficult. In contrast, the transferability $f_{X\rightarrow Y}$ is highly accurate between studies where virus behavior can be consistently extrapolated [Table 1, Datasets$_{Vac,1/2}$ and Datasets$_{Vac,3/4}$]. Moreover, as we show in the following section, the estimated error becomes more precise when we use multiple datasets to infer virus behavior.

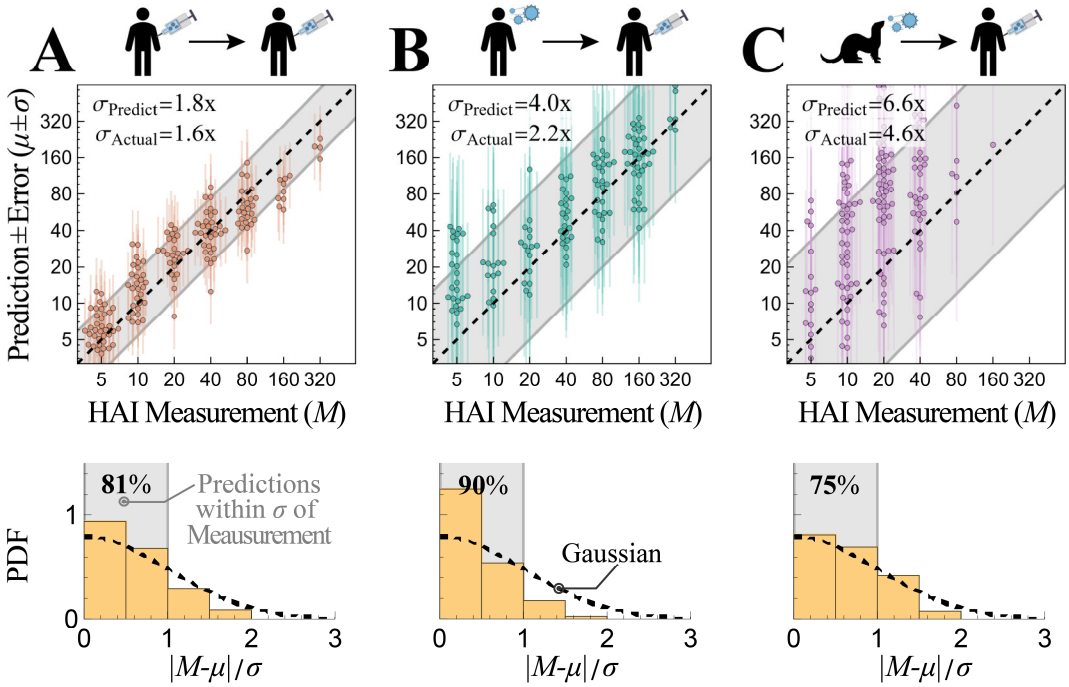

Figure 2. Predicting virus behavior between two datasets. Example predictions between two Fonville studies. *Top*, Plot comparing predicted and withheld HAI measurements (which take the discrete values 5, 10, 20…). Estimated error is shown in two ways: (*i*) as vertical lines emanating from each point and (*ii*) by the diagonal gray bands showing the average $\sigma_{Predict}$ of all predictions. *Bottom*, Histogram of the standardized absolute prediction errors compared to a standard folded Gaussian

distribution [black dashed line]. The fraction of predictions within 1.0σ are shown in the top left, which can be compared with the expected 68% for the standard folded Gaussian distribution. (A) Predicting A/Auckland/5/1996 between two human vaccination studies ($\text{Datasets}_{\text{Vac,3}\rightarrow\text{Vac,4}}$). (B) Predicting A/Netherlands/620/1989 between a human infection and human vaccination study ($\text{Datasets}_{\text{Infect,1}\rightarrow\text{Vac,4}}$). (C) Predicting A/Perth/27/2007 between a ferret infection and human vaccination study ($\text{Datasets}_{\text{Ferret}\rightarrow\text{Vac,4}}$).

Combining Influenza Datasets to Predict 200,000 Measurements with ≤3-fold Error

When multiple datasets are available to predict virus behavior in Dataset 1, we obtain *predictions±errors* ($\mu_j \pm \sigma_j$) from Dataset 2→1, Dataset 3→1, Dataset 4→1… These predictions and their errors are combined using the standard approach yielding

$$\frac{\Sigma_j\,(\mu_j/\sigma_j^2)}{\Sigma_j\,(1/\sigma_j^2)} \pm \frac{1}{[\Sigma_j(1/\sigma_j^2)]^{1/2}}. \qquad \text{(Equation 1)}$$

The uncertainty term in this combined prediction has two key features. First, adding an additional dataset (with predictions $\mu_k \pm \sigma_k$) only decreases the uncertainty. Second, if a highly uninformative dataset is added (with $\sigma_k \rightarrow \infty$), it will negligibly affect the cumulative prediction. Therefore, as long as the uncertainty estimates are reasonably precise, datasets do not need to be prescreened before matrix completion, and adding more datasets will always result in lower uncertainty.

To test the accuracy of combining multiple datasets, we performed leave-one-out analysis using all six Fonville studies, systematically withholding every virus in each dataset (311 virus-dataset pairs) and predicting the withheld values using all remaining data. Each dataset analyzed between 35-300 sera against 20-75 viruses (with 81 unique viruses across all six studies) and contained 0.5-40% missing values [Figure 3A].

Across all datasets, we predicted these 50,000 measurements with a low error of σ_{Actual}=2.1-fold. Upon stratifying these predictions by dataset, we found that the four human vaccinations studies were predicted with the highest accuracy ($\text{Datasets}_{\text{Vac,1-4}}$, σ_{Actual}≈2-fold) while the human infection study had slightly lower accuracy ($\text{Dataset}_{\text{Infect,1}}$, σ_{Actual}=2.6-fold) [Figure 3A]. Remarkably, even the least accurate human→ferret predictions had ≤4-fold error on average (σ_{Actual}=3.7-fold), demonstrating the potential for these cross-study inferences.

In addition to accurately predicting these values, the estimated error closely matched the true error in every human study ($\sigma_{\text{Predict}} \approx \sigma_{\text{Actual}}$, $\text{Datasets}_{\text{Vac,1-4}}$ and $\text{Dataset}_{\text{Infect,1}}$). In contrast, the uncertainty of the ferret predictions was highly overestimated (σ_{Predict}=6.4-fold, $\text{Dataset}_{\text{Ferret}}$) because none of the other datasets showed appreciable transferability to the ferret data. Said another way, the relationships between some viruses were the same in the human and ferret datasets, whereas other viruses showed different relationships across these two contexts. Given this highly-variable behavior, our framework overestimates σ_{Predict} to ensure that predictions with low uncertainty can always be trusted. Mathematically, this large estimated uncertainty arises because training error correlates poorly with prediction error, which results in steep transferability functions for the ferret data [Figure S1].

We visualize the transferability between datasets using a chord diagram [Figure 3B], where wider bands connecting $\text{Datasets}_{X\leftrightarrow Y}$ represent larger transferability [Figure S3, Methods]. As expected, there was high transferability between the human vaccine studies carried out in consecutive years

($Datasets_{Vac,1 \leftrightarrow Vac,2}$ and $Datasets_{Vac,3 \leftrightarrow Vac,4}$, Table 1), but far less transferability across vaccine studies more than 10 years apart ($Datasets_{Vac,1 \leftrightarrow Vac,3}$, $Datasets_{Vac,1 \leftrightarrow Vac,4}$, $Datasets_{Vac,2 \leftrightarrow Vac,3}$, or $Datasets_{Vac,2 \leftrightarrow Vac,4}$).

Transferability is not necessarily symmetric, since virus inhibition in Dataset *X* could exhibit all patterns in Dataset *Y* (leading to high transferability from $X \rightarrow Y$) along with unique patterns not seen in Dataset *Y* (resulting in low transferability from $Y \rightarrow X$). For example, none of the datasets showed appreciable transferability to the ferret dataset, although the ferret data could somewhat predict $Datasets_{Vac,3/4}$, and $Dataset_{Infect,1}$; this suggests the ferret data shows some patterns present in the human data but also includes additional unique structure. As another example, the human infection study carried out from 2007-2012 had high transferability from the human vaccine studies conducted in 2009 and 2010 ($Dataset_{Vac,3/4 \rightarrow Infect,1}$) but showed smaller transferability in the reverse direction.

While these results lay the foundation to compare different datasets, they are not exhaustive characterizations of each type of study – for example other human datasets using other viruses or sera may be able to predict these ferret responses more accurately. The strength of this approach lies in the fact that cross-study relationships are learned in a data-driven manner. In particular, as more datasets are added, the number of predictions between datasets increases while the uncertainty of these predictions decreases.

Breadth of Matrix Completion: Predicting Values based on a Distinct HAI Assay using only 5 Overlapping Viruses

To test the limits of our approach, we used the Fonville datasets to predict values from a large-scale serological dataset by Vinh *et al*. where *only 6* influenza viruses were measured against 25,000 sera (Vinh et al., 2021). This exceptionally long-and-skinny matrix is challenging for several reasons. First, after entirely withholding a virus, there are only 5 other viruses that can be used to infer its behavior. Furthermore, only 4/6 of the Vinh viruses had exact matches in the Fonville dataset; to utilize the remaining 2 viruses, we associated them with the closest Fonville virus based on their hemagglutinin sequences [Methods].

Second, the Vinh study measured HAI using protein microarrays (resulting in a continuum of values rather than the discrete 2-fold increments in Fonville), and such differences in design can lead to systematic shifts in the data.

Third, there were only 1,200 sera across all Fonville datasets, and hence predicting the behavior of 25,000 Vinh sera will be impossible if they all exhibit distinct phenotypes. Indeed, any such predictions would only be possible if the behavior of these 6 viruses is highly correlated (in both Fonville and Vinh), and if this multitude of serum responses is very degenerate.

Finally, even if accurate predictions are possible, it is hard to predict *a priori* which of the Fonville datasets would be most informative. The Vinh human infection study ($Dataset_{Infect,2}$) was carried out in Vietnam from 2009-2015, suggesting that either the Fonville human infection study ($Dataset_{Infect,1}$, carried out in Vietnam during similar years) or vaccine studies ($Datasets_{Vac,3/4}$ from 2009-2010) would be the best candidates [Table 1]. Surprisingly, we found that the greatest contribution came from an

earlier Fonville vaccine study ($\text{Dataset}_{\text{Vac},2}$ from 1998) and the human infection study [Figure 3B]. ($\text{Datasets}_{\text{Vac},3/4}$ could not make any predictions because they only contained 1/6 of the Vinh viruses.)

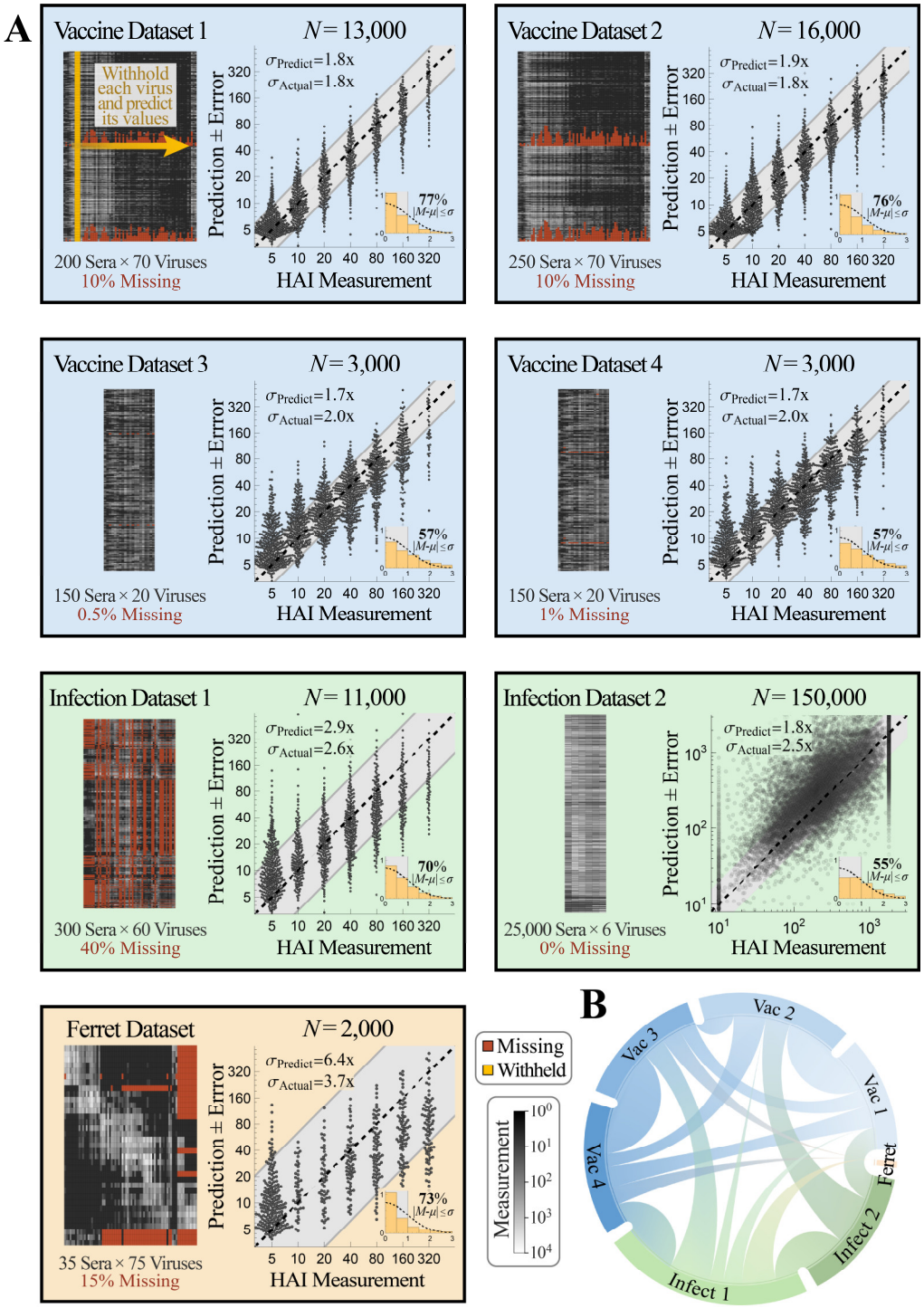

Figure 3. Validating predictions and error quantification across 200,000 measurements. (A) We combined seven influenza datasets spanning human vaccination studies (blue boxes), human infection studies (green), and a ferret infection study (orange). Each virus in every dataset was withheld and predicted using the remaining data (shown schematically in gold within the top-left box). For each dataset, we show a schematic of the full data (*left*, missing values in dark-red and measurements in

grayscale) together with a plot of the cumulative predictions for all viruses in that dataset (*right*, gray diagonal bands show the average predicted error $\sigma_{Predict}$). The total number of predictions *N* from each dataset is shown above its scatterplot; when this number of points is too great to show, we subsampled each distribution evenly while maintaining its shape. The inset at the bottom-right of each plot shows the PDF histogram of error measurements [*y*-axis] that were within 0.5σ, 1.0σ, 1.5σ… [*x*-axis], compared to a standard folded Gaussian distribution (black curve). The fraction of predictions within 1.0σ is explicitly written, which can be compared with the expected 68% for a standard folded Gaussian. (B) Chord diagram representing the transferability between datasets. For each arc connecting datasets *X* and *Y*, a larger width at the outer circle represents greater transferability [see Figure S3 and Methods].

After growing a forest of decision trees to establish the transferability between the Fonville and Vinh datasets [Figure S1], we predicted the 25,000 serum measurements for all 6 Vinh viruses with an average σ_{Actual}=2.5-fold error, demonstrating that even a small panel containing 5 viruses can be expanded to predict the behavior of additional strains [Figure 3, $Dataset_{Infect,2}$].

Notably, 5/6 of these viruses (which all circulated between 2003-2011) had a very low $\sigma_{Predict}\approx\sigma_{Actual}\approx$2-fold error [Figure S4]. The final Vinh virus circulated three decades earlier (in 1968), and its larger prediction error was underestimated (σ_{Actual}=4.5-fold, $\sigma_{Predict}$=2.7-fold). This inaccurate uncertainty highlights a shortcoming of any matrix completion algorithm, namely, that when a dataset contains one exceptionally distinct column (*i.e.*, one virus circulating 30 years before all other viruses), its values will not be accurately predicted. Nevertheless, we note that even predicting values with 4.5-fold error would save tens of thousands of experiments.

Leave-Multi-Out: Designing a Minimal Virus Panel that Maximizes the Information Gained per Experiment

Given the accuracy of leave-one-out analysis and that only 5 viruses are needed to expand a dataset, we reasoned that most studies include many viruses whose behavior could have been inferred. Pushing this to the extreme, we combined the Fonville and Vinh datasets and performed *leave-multi-out analysis*, where multiple viruses were simultaneously withheld and recovered, searching for a "minimal virus panel" that could recover each dataset. Future studies seeking to measure any set of viruses V_1-V_n can use a similar approach to select the smallest subset that can predict the full panel.

In the present search, we sought a minimal virus panel that recovers all Fonville and Vinh measurements with ≤4-fold error for each withheld virus. A virus was randomly selected from a dataset and added to the withheld list if its values, and those of all other withheld viruses, could be predicted with $\sigma_{Predict}$≤4-fold (without using σ_{Actual} to confirm these predictions, Methods). In this way, 124 viruses were concurrently withheld, representing 25–60% of the virus panels from every dataset or a total of *N*=70,000 measurements [Figure 4A].

Even with this hefty withheld set, the predictions only exhibited slightly larger errors than during leave-one-out analysis (σ_{Actual} between 2.1–3.1-fold for the human datasets and σ_{Actual} =3.8-fold for the ferret data). This small increase is due to two competing factors. On the one hand, prediction is far harder with fewer viruses. Yet at the same time, we specifically withheld highly-predictable viruses with $\sigma_{Predict}$≤4-fold (*i.e.*, each withheld strain is well-characterized by other viruses). These mostly factors offset one another, so that the 70,000 measurements exhibited the desired σ_{Actual}≤4-fold.

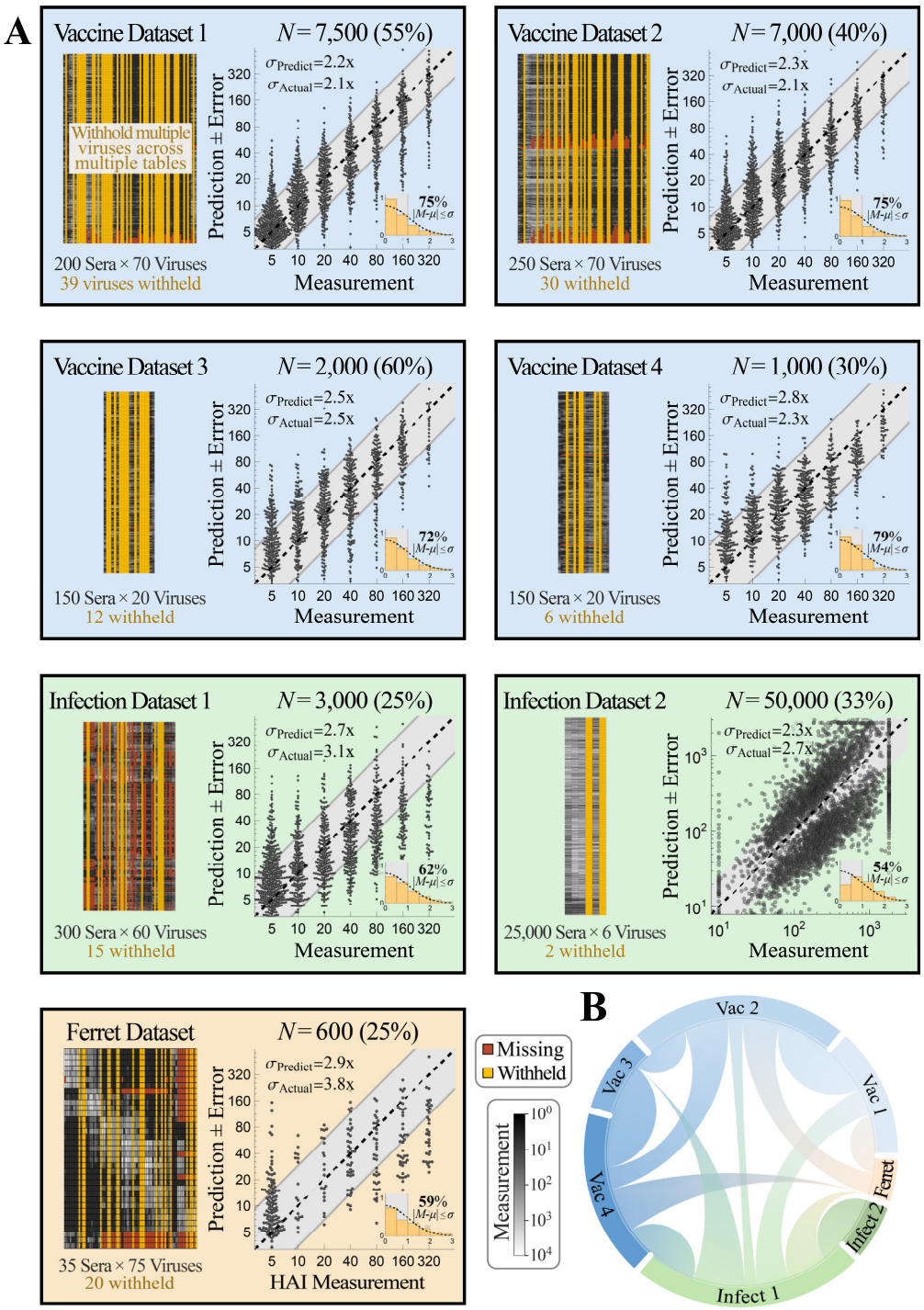

Figure 4. Simultaneously predicting measurements for 124 viruses withheld from multiple datasets. (A) Viruses were concurrently withheld from each dataset (gold rows in dataset schematics), and their 70,000 values were predicted using the remaining data. We sought to withhold as many viruses as possible while still retaining a low estimated error ($\sigma_{Predict} \leq$4-fold), and indeed, the actual prediction error was smaller than 4-fold in each dataset. As in Figure 3, plots and histograms show the collective predictions and error distributions in each dataset. The plot label shows the number of concurrent predictions (and percent of data predicted). (B) Chord diagram representing the transferability between datasets after withholding the viruses. For each arc connecting datasets *X* and *Y*, a larger width at the outer circle represents greater transferability [see Figure S3 and Methods].

The transferability between datasets, computed without the withheld viruses, was similar to the transferability between the full datasets [Figure 4B]. Some connections were lost when there were too few overlapping viruses between datasets, while other connections were strengthened when the patterns in the remaining data became more similar across studies. Notably, the ferret data now showed some transferability from vaccination Datasets$_{Vac,1/2}$, which resulted in smaller estimated error than in our leave-one-out analysis ($\sigma_{Predict}$=2.9-fold). This emphasizes that transferability depends upon the specific viruses and sera examined, and that some parts of the Fonville human dataset can inform the ferret data. While this uncertainty underestimated the true error of the ferret predictions (σ_{Actual}=3.8-fold), they remained within the desired 4-fold error threshold. Moreover, in all six human datasets, the estimated uncertainty $\sigma_{Predict}$ closely matched the true error σ_{Actual}, demonstrating that there is significant potential to predict the behavior of many new viruses within each dataset.

Extending all Datasets to include $2{\cdot}10^6$ New Measurements with a Single, Unified Virus Panel

In the previous section, we combined datasets to predict the behavior of additional viruses, validating our approach on 200,000 existing measurements. Future studies can immediately leverage the Fonville datasets to expedite their efforts. If a new dataset contains at least 5 Fonville viruses [green arrows/boxes in Figure 5A], the values±errors for the remaining Fonville viruses can be predicted. Viruses with an acceptably low error [purple in Figure 5A] can be added without requiring any additional experiments.

To demonstrate this process, we first focus on the Vinh dataset where expansion will have the largest impact, since the Vinh virus panel is small (6 viruses) but its serum panel is enormous (25,000 sera). By predicting the interactions between these sera and all 81 unique Fonville viruses, we add 2,000,000 new predictions (more than 10x the number of measurements in the original dataset).

For each Fonville virus V_0 that was not measured in the Vinh dataset, we grew a forest of decision trees as described above, with the minor modification that the 5 features must be restricted to the Vinh virus panel to enable its expansion. The top trees were combined with the transferability functions (shown in Figure S1) to predict the values±errors for V_0.

The majority of the added Fonville viruses (68/75) were predicted with a low uncertainty of $\sigma_{Predict}$≤4-fold [Figure 5B]. As expected, viruses circulating around the same time as the Vinh panel (1968 or 2003-2011) tended to have the lowest uncertainty, whereas viruses from the 1990s had the largest uncertainty [Figure 5C]. To confirm these estimates, we restricted the Fonville datasets to these same 6 viruses and expanded out, finding that any virus with $\sigma_{Predict}$≤6-fold prediction error (as is the case for all Vinh predictions) had a true error σ_{Actual}≤6-fold [Figure S5].

The power of this expansion lies not only in the increased resolution of these antibody responses, but also in the ability to directly compare responses using a unified virus panel. For example, only 9/81 viruses were measured in all six Fonville datasets, hampering cross-study comparisons. After repeating the above expansion process for the Fonville datasets (adding 175 new columns across the six studies), every antibody response is predicted against all 81 viruses. Even after excluding the 6000/74000 (8%) of predictions with estimated error $\sigma_{Predict}$>4-fold, this leaves ample opportunities to

compare the Fonville and Vinh data. These extended datasets are provided in the Github repository associated with this paper.

As additional datasets are added, both the Fonville and Vinh datasets can be further expanded, and they will expand the added datasets in return. For each new dataset, we incur a one-time computational cost to grow a forest of decision trees to quantify its transferability, after which predicting new virus behavior takes a fraction of a second. This scalable approach is well-suited to the varied landscape of serological studies where experimental design (*e.g.*, the organism studied, assay employed, or virus panel utilized) can change between studies, and a quantitative and data-driven approach will be highly effective.

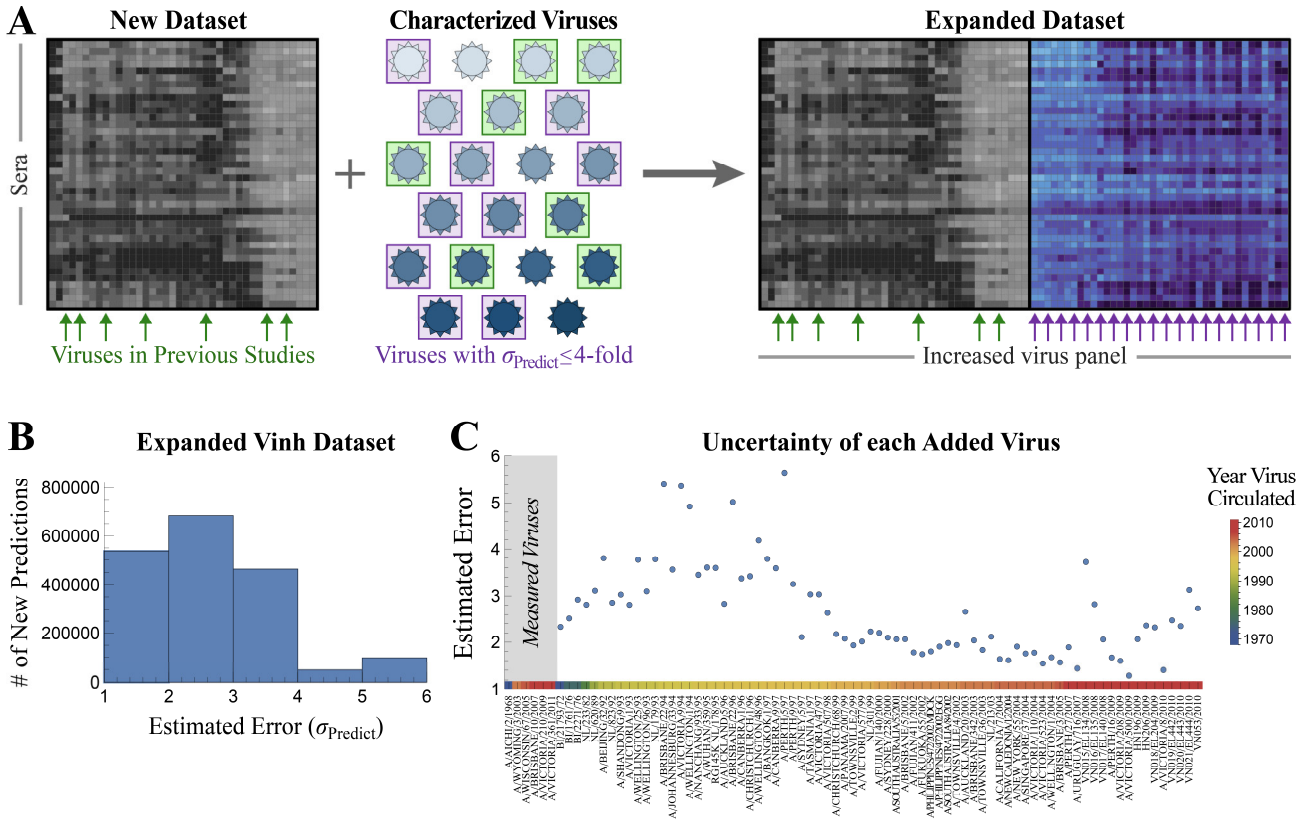

Figure 5. Expanding the Vinh dataset with 75 additional viruses. (A) If a new study contains at least 5 previously characterized viruses [green boxes and arrows], we can predict the behavior of all previously characterized viruses in the new dataset. Those with an acceptable error (*e.g.*, ≤4-fold error boxed in purple) are used to expand the dataset. (B) Predicting how each Fonville virus will inhibit the 25,000 Vinh sera. Distribution of the estimated uncertainty $\sigma_{Predict}$ for these additional measurements. The majority of viruses are estimated with ≤4-fold error. (C) Estimated uncertainty for each added virus. The six viruses on the left represent the Vinh virus panel. Colors at the bottom represent the year each virus circulated.

Matrix Completion via Nuclear Norm Minimization Poorly Predicts Behavior across Studies

In this final section, we briefly contrast our algorithm against singular value decomposition (SVD) based approaches such as nuclear norm minimization (NNM), which are arguably the simplest and best-studied matrix completion methods. With NNM, missing values are filled by minimizing the sum of singular values of the completed dataset.

To compare our results, we reran our leave-multi-out analysis from Figure 4, simultaneously withholding 124 viruses and predicting their values using an established NNM algorithm from (Einav and Cleary, 2022). The predictions in each table were markedly worse, with the Fonville and Vinh datasets exhibiting σ_{Actual} between 3.4x-5.4x.

Due to two often-neglected features of NNM, we find that our approach significantly outperforms this traditional route of matrix completion in predicting values for a completely withheld virus column. First, one artifact of NNM is that there is an asymmetry when predicting large and small values for the withheld virus. Consider a simple noise-free example where one virus's measurements are proportional to another's, (Virus 2's values)=$m\cdot$(Virus 1's values) with m=5 [Figure S6A]. Surprisingly, if measurements of Virus 2 are withheld, NNM incorrectly predicts that (Virus 2's values)=(Virus 1's values) for any $m\geq 1$ [Figure S6B]. This behavior is exacerbated when multiple datasets are combined, emphasizing that NNM can catastrophically fail for very simple examples [Figure S6C,D]. This artifact can be alleviated by first row-centering a dataset before matrix completion, as in Algorithm 1.

Yet even with row-centering, a second artifact of nuclear norm minimization is that large swaths of missing values can skew matrix completion when patterns are (incorrectly) inferred between the missing values. Intuitively, all iterative NNM algorithms must initialize the missing entries (often either with 0 or the row/column means), so that two viruses with very different behaviors may end up appearing identical across their missing values. For example, suppose we want to predict values for virus V_0 from Dataset $X\rightarrow Y$, and that "useful" viruses V_1-V_4 behave similarly to V_0 in Datasets X and Y. On the other hand, "useless" viruses V_5-V_8 are either not measured in Dataset 2 or are measured against complementary sera, and moreover these viruses show very different behavior from V_0 in Dataset 1 [Figure S7 shows a concrete example from Fonville]. Ideally, a matrix completion algorithm should ignore V_5-V_8, given that they do not match V_0 in Dataset 2, and only use V_1-V_4 to infer V_0's values in Dataset 1. In practice, matrix completion using V_0-V_8 results in poor predictions [Figure S7]. This behavior is disastrous for large serological datasets, where there can be >50% missing values when datasets are combined.

Our algorithm was constructed to specifically avoid both of these artifacts. First, we infer each virus's behavior using a decision tree on row-centered data which does not exhibit the asymmetry discussed above. Second, we restrict our analysis to features that have ≥80% observed measurements to ensure that patterns detected are based on measurements rather than on missing data.

As another point of comparison, consider the leave-one-out predictions of the six Vinh viruses using the Fonville datasets. Whereas our algorithm yields tight predictions that span the full range of measured HAI values (Figure S4), NNM led to a nearly flat response with all 25,000 sera incorrectly predicted to be the mean of the measurements [see Figure S8 in (Einav and Cleary, 2022)]. Finally, we emphasize that there is currently no robust method to compute the error of NNM when predicting new virus behavior across datasets, and that error quantification is essential to understand which predictions can be trusted.

Discussion

As experiments voraciously grow in size and scope, it is easy to overlook that the wealth of available data can greatly expedite future work. Here, we developed an algorithm that leverages patterns in antibody-virus data to predict how a virus measured in one study would behave in another study, without requiring any additional experiments.

A major hurdle in cross-study inferences is reproducibility. Many anecdotes and formal studies have demonstrated that subtle differences in a protocol can lead to vastly different results (Hines et al., 2014; Zacour et al., 2016). Our algorithm enables cross-study predictions in two key ways. First, we determine the transferability between two studies using their overlapping viruses, quantifying how accurately the behaviors in one study map to the other (Figure 3B, Figure S1). Using this transferability, we can estimate the uncertainty of each prediction without requiring information about study design or infection history. Second, we account for systematic shifts between datasets by developing a random forest algorithm using row-centered data (Methods), making our predictions far more robust. As a result, we inferred virus behavior between the Fonville and Vinh studies, even though they utilized different methods of HAI, had different dynamic ranges for their data, and used markedly different virus panels (Fonville et al., 2014; Vinh et al., 2021). After validating our predictions on available data, we expanded the Vinh panel to include 75 new viruses with $2 \cdot 10^6$ new measurements, increasing the original dataset by >10x. Similar expansions can and should be carried out for future studies.

Our results suggest that instead of thinking about each serum sample as being entirely unique, large collections of sera often exhibit surprisingly similar inhibition profiles. For example, the 1,200 sera in the Fonville datasets predicted the behavior of the 25,000 Vinh sera with ≤2.5-fold error on average, demonstrating that these Vinh sera were *at least* 20-fold degenerate. As studies continue to probe the diverse pool of sera worldwide, their transferability will determine when new sera exhibit fundamentally different behavior.

The transferability also provides a succinct way to quantify the similarity between two datasets. Since transferability is solely computed from the data, we can layer the details of each study [*e.g.*, participants' age, geographic location, infection history (Lewnard and Cobey, 2018; Henry et al., 2019; Dugan et al., 2020)] on top of these results. In this work, we found surprisingly large transferability between human infection and vaccination studies. For example, vaccine studies from 1997/1998 ($\text{Dataset}_{\text{Vac,1/2}}$) were highly informed by the Vinh infection study from 2009-2015 ($\text{Dataset}_{\text{Infect,2}}$), even though none of the Vinh participants had ever been vaccinated. Conversely, both infection studies we analyzed were highly informed by vaccine studies (*e.g.*, $\text{Dataset}_{\text{Infect,1}}$ was most informed by $\text{Datasets}_{\text{Vac,3/4}}$). These results demonstrate that even studies with very different experimental setups can be combined to predict new virus behavior.

The decision trees underlying our random forest algorithm are highly modular, and additional trees can be grown to carry out a specific analysis. For example, in addition to the forest that powered our leave-one-out analysis, we grew another forest when simultaneously withholding 124 viruses to gain extra resolution, and these forests were combined in our leave-multi-out analysis (Methods). Similarly, adding a new dataset does not require remaking existing forests, but rather entails creating additional trees

that link the new study to existing datasets. Through this approach, the relationships between every pair of studies weaves together to form a vast global network.

Future studies should capitalize on existing datasets, saving time and effort by choosing smaller virus panels that can be subsequently expanded (Figure 5A). One powerful approach is to perform experiments in stages, interspersing matrix completion to inform next steps. For example, a new study assessing serum inhibition against viruses V_1-V_n should first determine the minimal subset of viruses that can predict the full panel in existing datasets. After measuring this subset of viruses, the transferability to this new dataset can be computed and the values±errors across the full panel can be estimated. Importantly, at this stage our framework also predicts how measuring any additional virus would further decrease the estimated uncertainties (since predictions are based upon virus relationships in *existing* datasets). In other words, this approach predicts which additional experiments will be maximally informative to achieve tight predictions.

Matrix completion draws upon the collective information gained in multiple studies to infer how an antibody response would inhibit a broad array of viruses. Such efforts paint the immunological landscape in far greater resolution, fueling diverse research efforts ranging from viral surveillance (Morris et al., 2018) to characterizing the composition of antibodies within serum (Georgiev et al., 2013; Fonville et al., 2014; Lee et al., 2019a; Einav et al., 2022) to predicting future antibody-virus coevolution (Sheng and Wang, 2021; Marchi et al., 2021). Although we focused on influenza HAI data, this method can be readily applied to other inherently low-dimensional datasets, both in and out of immunology. In the context of antibody-virus interactions, this approach not only massively extends current datasets, but also provides a level playing field where antibody responses from different studies can be directly compared using the same set of viruses. This shift in perspective expands the scope and utility of each measurement, enabling future studies to always build on top of previous results.

Methods

Availability of Code and Results

The code to perform matrix completion in *Mathematica* and *R* are included in the associated GitHub (https://github.com/TalEinav/CrossStudyCompletion), which also contains the expanded Fonville and Vinh datasets.

Datasets Analyzed

Information about the Fonville and Vinh datasets (type of study, year conducted, and geographic location) is provided in Table 1. The number of sera, viruses, and missing measurements in each dataset is listed below the schematics in Figure 3.

Every serum was unique, appearing in a single study. All Fonville viruses appeared in at least 2 studies, enabling us to entirely remove a virus from one dataset and infer its behavior from another dataset.

Although the Vinh dataset contained H1N1 and H3N2 viruses, we only considered the 6 H3N2 strains (since Fonville only contained H3N2 viruses). 4/6 of the Vinh viruses (H3N2 A/WYOMING/3/2003, A/WISCONSIN/67/2005, A/BRISBANE/10/2007, and A/VICTORIA/361/2011) were found in the Fonville

datasets. The remaining 2 viruses were not in Fonville, and hence we associated each of them with the Fonville virus that had the most similar HA sequence (Vinh virus A/AICHI/2/1968→Fonville virus BI/16190/68; Vinh virus A/VICTORIA/210/2009→Fonville virus HN201/2009). While such substitutions may increase prediction error (which can be gauged through leave-one-out analysis), they also vastly increase the number of possible cross-study predictions, since two virus panels only need to have homologous viruses rather than exact matches.

Influenza Dataset	Organism	Type	Year Conducted	Geographic Location	Source of Data
$\text{Dataset}_{\text{Vac},1}$	Human	Vaccination	1997	Parkville, Australia	Fonville 2014, Table S5
$\text{Dataset}_{\text{Vac},2}$	Human	Vaccination	1998	Parkville, Australia	Fonville 2014, Table S6
$\text{Dataset}_{\text{Vac},3}$	Human	Vaccination	2009	Parkville, Australia	Fonville 2014, Table S13
$\text{Dataset}_{\text{Vac},4}$	Human	Vaccination	2010	Parkville, Australia	Fonville 2014, Table S14
$\text{Dataset}_{\text{Infect},1}$[(b)]	Human	Infection	2007-2012	Ha Nam, Vietnam	Fonville 2014, Table S3
$\text{Dataset}_{\text{Infect},2}$[(b)]	Human	Infection	2009-2015	Ha Nam, Vietnam	Vinh 2021 Supplement
$\text{Dataset}_{\text{Ferret}}$[(a)]	Ferret	Infection	N/A	N/A	Fonville 2014, Table S1

Table 1. Datasets analyzed in this work. Information about the type of study as well as the year and geographic location from which the antibody responses were collected.
[(a)] Infected influenza-naive ferrets with a single virus and measured their serum against a panel of viruses.
[(b)] Over multiple years, participants reported influenza-like illnesses and got PCR tested. Serum samples were collected from all participants once each year.

Matrix Completion on $\log_{10}$(titers)

As in previous studies, all analysis was done on $\log_{10}$(HAI titers) because experimental measurements span orders of magnitude, and we did not want to bias the predictions towards the largest values (Einav and Cleary, 2022). Thus, when computing the distribution of errors (histogram in Figures 2-4), each of M, μ, and σ are computed in $\log_{10}$. Anytime we discussed a prediction or its error, we exponentiated by 10 (*i.e.*, $\sigma_{\text{Predict,log10}}=0.3$ for $\log_{10}$ titers corresponds to an error of $\sigma_{\text{Predict}}=10^{0.3}=2$-fold).

In the Fonville dataset, we replaced lower or upper bounds by their next 2-fold increment ("<10"→5 and "≥1280"→2560). The Vinh dataset did not include any bounded measurements, although their HAI titers exhibit clear signs of being clipped between 10 and 1810, as can be seen by plotting the values of any two viruses across all sera. This is why the Vinh predictions in Figure 3 ($\text{Dataset}_{\text{Infect},2}$) contains multiple points on the left and right edges of the plot.

Error of the Hemagglutination (HAI) Assay

In (Fonville et al., 2014), analysis of repeated HAI measurements showed that the inherent error of the assay is log-normally-distributed with standard deviation $\sigma_{\text{Inherent}}\approx 2$-fold. This is shown by Figure S8B of (Fonville et al., 2014) [neglecting the stack of not-determined measurements outside the dynamic range

of the assay, which also must be neglected from their Figure S8A], where 40% of repeats had the same HAI value, 50% had a 2-fold discrepancy, and 10% had a 4-fold discrepancy.

Using Decision Trees to Quantify the Relationships between Viruses

Decision trees are a simple, easily-interpretable, and well-studied form of machine learning. An advantage of decision trees is that they are very fast to train, with implementations in many programming languages. The predictions from decision trees are made even more robust using a random forest approach, which in our case involved averaging the results from 5 top trees.

As described in Algorithm 1, we trained regression trees that take as input the row-centered $\log_{10}$(HAI titers) from viruses V_1-V_5 to predict another virus V_0. These trees can then be applied in another dataset to predict V_0 based on the values of V_1-V_5.

Row-centering means that if we denote the $\log_{10}$(titers) of V_0-V_5 to be t_0-t_5 with mean t_{avg}, then the decision tree will take as input (t_1-t_{avg}, t_2-t_{avg}, t_3-t_{avg}, t_4-t_{avg}, t_5-t_{avg}) and attempt to predict t_0-t_{avg}. The value t_{avg} (which will be different for each serum) is then added to the output of the decision tree to undo the row-centering. If any of the t_j are missing (including t_0 when we withhold V_0's values), we proceed in the same way but compute t_{avg} as the average of the measured values. Row-centering enables the algorithm to handle systematic differences in data. For example, the analysis is independent of which units are chosen for the data (*e.g.*, neutralization measurements in µg/mL or Molar would both be handled the same, since in $\log_{10}$ they are offset from each other by a constant factor that will be subtracted during row-centering). If one serum is concentrated by 2x (or any other factor), its titers would all increase by 2x but the relationships between viruses would remain the same; row-centering subtracts this extra concentration factor and yields the same analysis.

When training our decision trees, we allow missing values for V_1-V_5 (but not V_0) [as shown by the schematic in Figure 1B], with these missing values replaced by the most likely value (*i.e.,* mode-finding) given the known values in the training set. When applying a trained decision tree to other datasets, we only predicted a value for V_0 when none of V_1-V_5 were missing (otherwise that decision tree was ignored). If all 5 top trees were ignored due to missing values, then no prediction was made for this interaction of V_0 and serum, and this value was ignored in subsequent analyses.

Predicting the Behavior of a New Virus

As described in Algorithm 1, the values for V_0 predicted from Dataset $D_j \rightarrow D_0$ is based on the top 5 decision trees that predict V_0 in D_j with the lowest σ_{Training}. The value of V_0 against any serum is given by the average value of the top 5 decision trees, while its error is given by the estimated error $\sigma_{\text{Predict}}=f_{D_j \rightarrow D_0}(\sigma_{\text{Training}})$ of these top 5 trees, where $f_{D_j \rightarrow D_0}$ represents the transferability map [described in the next section]. Thus, every prediction of V_0 in D_0 will have the same σ_{Predict}, unless some of the top 5 trees cannot cast a vote because their required input titers are missing (in which case the value±error is computed as the average of the top 5 trees that can vote). In practice, the estimated error for V_0 in D_0 is overwhelmingly the same across all sera, as seen in Figure 2 where the individual error of each measurement is shown via error bars.

The estimated error σ_{Predict} and true error σ_{Actual} in Figures 3 and 4 were computed using all data. When scatter plots contained too many data points to show with appreciable resolution, we subsampled each distribution evenly across its predicted value to maintain its shape. We did not display the small fraction of measurements with HAI titers≥640 in order to better show the portions of the plots with the most points; however, error statistics were computed using all data.

<u>Algorithm 1</u>: Predicting Virus Behavior (Value±Error) across Studies via a Random Forest

Input:

- Dataset-of-interest D_0 containing virus-of-interest V_0 whose measurements we predict;
- Other datasets $\{D_j\}$, each containing V_0 and at least 5 viruses $V_{j,1}$, $V_{j,2}$… that overlap with the D_0 virus panel, used to extrapolate virus behavior;
- Antibody responses $A_{j,1}$, $A_{j,2}$… in each dataset D_j. When $j \neq 0$, we only consider the antibody responses with non-missing values against V_0

Steps:

1. For each D_j, create n_{Tree}=50 decision trees predicting V_0 based on n_{Features}=5 other viruses and a fraction f_{Samples}=3/10 of sera
    - For robust training, we restrict attention to features with ≥80% non-missing values. If fewer than n_{Features} viruses in D_j satisfy this criterion, do not grow any decision trees for this dataset
    - Bootstrap sample (with replacement) both the viruses and antibody responses
    - Data is analyzed in $\log_{10}$ and row-centered on the features (*i.e.*, for each antibody response in both the training set D_j and testing set D_0, subtract the mean of the $\log_{10}$[titers] for the n_{Features} viruses using all non-missing measurements). This row-centering accounts for systematic shifts between datasets. Once the decision tree makes its predictions, this row-centering is then undone (by adding this serum-dependent mean)
    - Compute the cross-validation RMSE (σ_{Training}) of each tree against the unused fraction 1-f_{Samples} of samples in D_j
2. Use the $n_{\text{BestTrees}}$=5 decisions trees with the lowest σ_{Training} to predict the (un-row-centered) values of V_0 in D_0
    - Only make predictions for antibody responses in D_0 where *all* n_{Features} are non-missing. If any feature is missing, this tree makes no prediction for this antibody response
    - For each antibody response, we predict the response $\mu_j \pm \sigma_j$
        - μ_j=(mean value for the $n_{\text{BestTrees}}$ predictions)
        - $\sigma_j = f_{D_j \to D_0}$ (mean σ_{Training} for the $n_{\text{BestTrees}}$ trees), where the transferability $f_{D_j \to D_0}$ is computed by predicting $V_{j,1}$, $V_{j,2}$… in D_0 using D_j (see Algorithm 2)
3. Combine predictions for V_0 in D_0 using all other datasets $\{D_j\}$ using $\frac{\Sigma_j\,(\mu_j/\sigma_j^2)}{\Sigma_j\,(1/\sigma_j^2)} \pm \frac{1}{[\Sigma_j(1/\sigma_j^2)]^{1/2}}$

Transferability Map between Datasets

A key component of our analysis is to quantify how the error of a decision tree trained in dataset D_j translates into this tree's error in dataset D_0. Importantly, when predicting the behavior of a virus V_0 in D_0, we *cannot* access V_0's values and hence cannot directly compute the actual error of this tree.

To solve this problem, we ignore V_0 and apply Algorithm 1 to predict the values of many viruses that are in both D_0 and D_j. Since we can access the values of these viruses in both datasets, we can directly compare their σ_{Training} in D_j against σ_{Actual} in D_0. We did not know *a priori* what the relationship would be between these two quantities, and surprisingly, it turned out to be well-characterized by a simple linear relationship (Figure S1).

As described in Algorithm 2, we obtain the best-fit line to these data (and we found that using perpendicular offsets yields a more intuitive fit). We add a vertical shift to account for scatter about the best-fit line, thereby ensuring that in highly-variable cases where some trees have small σ_{Training} but large σ_{Actual} (*e.g.*, $\text{Dataset}_{\text{Ferret}} \rightarrow \text{Dataset}_{\text{Vac,1}}$), we tend to overestimate rather than underestimate the error. In other words, some predictions with large uncertainty may still be accurate, but all predictions with low uncertainty should be accurate.

To visualize the transferability maps between every pair of datasets, we construct a chord diagram where the arc connecting Dataset X and Y represents a double-sided arrow quantifying both the transferability from $X \rightarrow Y$ [thickness of the arc at Dataset Y] as well as the transferability from $Y \rightarrow X$ [thickness of the arc at Dataset X] (Figure S3). More precisely, the width of each arc is equal to $\Delta\theta \equiv (2\pi/16)(\partial f_{D_j \rightarrow D_0}/\partial\sigma_{\text{Training}})^{-1}$, so that width is proportional to 1/slope of the transferability best-fit line from Figure S1. We used the factor 16 in the denominator so that the chord diagrams in Figures 3B,4B would form nearly complete circles, and if more studies are added this denominator can be modified (increasing it would shrink all the arcs proportionally). Note that the size of the arcs in Figures 3B and 4B can be directly compared to one another, so that if the arc from $X \rightarrow Y$ is wider in one figure, it implies more transferability between these datasets. For cleanliness, we did not show the very weak transferability maps (*i.e.*, those with large slopes>5 in Figure S1), instead opting to only show the strong relationships between datasets.

We note that the transferability shown in Figure 3B and Figure S1 represent all antibody-virus data, which is slightly different from the transferability maps we use when predicting values for virus V_0 in dataset D_0. When withholding a virus, we made sure to remove from Figure S1 all trees using this virus. Although this can slightly change the best-fit line, in practice the difference is very minor. However, when withholding multiple viruses in our leave-multi-out analysis, the number of datapoints in Figure S1 substantially decreased, and to counter this we trained additional decision trees (as described in the following section).

<u>Algorithm 2</u>: Computing the Transferability $f_{D_j \to D_0}$ between Datasets

Input:

- Datasets $\{D_j\}$ that collectively include the viruses V_1, V_2… Each virus may be missing from some datasets, but must be included in at least two datasets

Steps:

- For each dataset $D_0 \in \{D_j\}$
  - For each virus V_0 in D_0
    - For every other dataset D_j containing V_0
      - Create n_{Tree}=50 decision trees predicting V based on n_{Features}=5 other viruses, as described in Algorithm 1
      - For each tree, store the following information:
        - D_0, V_0, and D_j used to construct the tree
        - Viruses used to train the tree
        - RMSE σ_{Training} on the 1-f_{Samples} samples in D_j
        - Predictions of V_0's values in D_0
        - True RMSE σ_{Actual} of these predictions for V_0 in D_0
- When predicting V_0 using $D_j \to D_0$ in Algorithm 1, we compute the relationship $f_{D_j \to D_0}$ between σ_{Training} and σ_{Actual} by predicting the other viruses V_1, V_2…V_n that overlap between D_j and D_0 (but excluding the withheld V_0)
  - From the forest of decision trees above, find the top 10 trees for each virus predicted between $D_j \to D_0$ and plot σ_{Training} vs σ_{Actual} for all trees [see Figure S1]. Ensure that the viruses used in these decision trees do *not* contain V_0
  - Find the best-fit line using perpendicular offsets, $y=ax+b$ where $x=\sigma_{\text{Training}}$ and $y=\sigma_{\text{Actual}}$. Since there is scatter about this best-fit line, and because it is better to overestimate rather than underestimate error, we add a correction factor c=(RMSE between the σ_{Actual} and $ax+b$). Lastly, we expect that a decision tree's error in another dataset will always be at least as large as its error on the training set ($\sigma_{\text{Actual}} \geq \sigma_{\text{Training}}$), and hence we define $f_{D_j \to D_0}$ = max($a\sigma_{\text{Training}}+b+c$, σ_{Training}). This max term is important in a few cases where $f_{D_j \to D_0}$ has a very steep slope but some decision trees have small σ_{Training}
  - Datasets with high transferability will have $f_{D_j \to D_0}(\sigma_{\text{Training}}) \approx \sigma_{\text{Training}}$, meaning that viruses can be removed from D_0 and accurately inferred from D_j. In contrast, two datasets with low transferability will have a nearly vertical line, $\partial f_{D_j \to D_0}/\partial\sigma_{\text{Training}} \gg 1$, signifying that viruses will be poorly predicted between these studies
  - In the chord diagrams (Figures 3B,4B), the width of the arc between Dataset D_j and D_0 is proportional to $(\partial f_{D_j \to D_0}/\partial\sigma_{\text{Training}})^{-1}$

Leave-Multi-Out Analysis

For this work, we trained many decision trees using different choices of viruses V_1-V_5 to predict V_0 in different datasets. For leave-one-out analysis, we created 50,000 trees, providing ample relationships between viruses. However, when we withheld 124 viruses during the leave-multi-out analysis, we were careful to not only exclude decision trees predicting one of these withheld viruses (as V_0), but to also

exclude decision trees using any withheld virus in the feature set (in V_1-V_5). As a result, only 6,000 trees out of our original forest exclusively used the non-withheld viruses, and this limited sampling of the relationships between viruses leads to worse σ_{Predict} and σ_{Actual} error. Thus, we grew additional trees specifically using these non-withheld viruses to make predictions. Once these trees are grown, we applied Algorithm 1 as before.

To find a minimal virus panel, we randomly choose one of the 317 virus-study pairs from the Fonville/Vinh datasets, adding it to the list of withheld viruses provided that each withheld entry could be predicted with ≤4-fold error. We note that given a forest of decision trees, it is extremely fast to test whether a set of viruses all have σ_{Predict}≤4-fold. However, as described above, as more viruses are withheld, our forest is trimmed which leads to poorer estimations of σ_{Predict}. As a result, we must grow more decision trees at some key steps.

Our procedure to find a minimal virus panel proceeded in three steps:

- *Step 1*: *Choose Vinh viruses to withhold, and then choose viruses from the Fonville human studies.* Because there are only 6 Vinh viruses, and removing any one of them from the Fonville datasets could preclude making any Vinh predictions, we first withheld 2 Vinh viruses. We then started withholding viruses from the human datasets (#1-4 and 6) where we had the most decision trees.
- *Step 2*: *Create an additional random forest for the Fonville ferret dataset (#5)*. Make sure to only use the non-withheld viruses from the other datasets as features (since any trees using those withheld viruses must be discarded). As described above, this additional forest quantifies the relationships between the remaining non-withheld viruses with higher resolution, enabling us to estimate σ_{Predict} more accurately. With this forest, choose additional viruses from the Fonville ferret dataset to withhold.
- *Step 3*: *Create additional random forests for the Fonville human datasets.* Use the improved resolution provided by these new forests to determine if any of the previously withheld viruses now have σ_{Predict}>4-fold and remove them. Finally, use the additional high-resolution forests to search for additional Fonville viruses to withhold.

Extending Virus Panels

When extending the Fonville and Vinh virus panels, we grew an additional forest of decision trees. Unlike in our leave-one-out analysis, the two key differences with this forest were that none of the data was withheld and that the allowed feature sets were specifically tailored to each dataset. For example, to expand the Vinh dataset to predict one of the 81-6=75 Fonville viruses V_0 (excluding the 6 viruses already in the Vinh data), we only searched for relationships between the six Vinh viruses and V_0 across the Fonville datasets. (Any decision trees using other viruses could not be applied to the Vinh dataset, which only contains 6 viruses.)

Once these additional trees were grown, we could predict the behavior of all 81 Fonville viruses against nearly every serum analyzed in the Fonville or Vinh datasets. The exceptions were sera such as those shown in the middle and bottom of $\text{Datasets}_{\text{Vac},1/2}$, which were measured against few viruses and hence we found no relationship between their available measurements in our random forest. The expanded virus panels are available in the GitHub repository associated with this paper.

Acknowledgements

We would like to thank Andrew Butler and Bernadeta Dadonaite for their input on this manuscript. Tal Einav is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRQ 01-20). Rong Ma is supported by Professor David Donoho at Stanford University.

Author Contributions

T.E. and R.M. conducted the research and wrote the paper.

Competing Interests

The authors declare no competing interests.

References

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

Candes, E.J., and Plan, Y. (2010). Matrix completion with noise. Proc. IEEE *98*, 925–936. https://doi.org/10.1109/JPROC.2009.2035722.

Candes, 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. Theory *56*, 2053–2080. https://doi.org/10.1109/TIT.2010.2044061.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. Proc. Natl. Acad. Sci. *116*, 22931–22937. https://doi.org/10.1073/PNAS.1910053116/-/DCSUPPLEMENTAL.

Dugan, H.L., Guthmiller, J.J., Arevalo, P., Huang, M., Chen, Y.Q., Neu, K.E., Henry, C., Zheng, N.Y., Lan, L.Y.L., Tepora, M.E., et al. (2020). Preexisting immunity shapes distinct antibody landscapes after influenza virus infection and vaccination in humans. Sci. Transl. Med. *12*, 3601. https://doi.org/10.1126/SCITRANSLMED.ABD3601.

Einav, T., and Cleary, B. (2022). Extrapolating Missing Antibody-Virus Measurements across Serological Studies. Cell Syst. (in Press. https://doi.org/10.1101/2021.08.29.458105.

Einav, T., Creanga, A., and Kanekiyo, M. (2022). Harnessing Low Dimensionality to Visualize the Antibody-Virus Landscape for Influenza. BioRxiv 2020.08.28.270561. https://doi.org/10.1101/2020.08.28.270561.

Fonville, J.M., Wilks, S.H., James, S.L., Fox, A., Ventresca, M., Aban, M., Xue, L., Jones, T.C., Le, N.M.H., Pham, Q.T., et al. (2014). Antibody landscapes after influenza virus infection or vaccination. Science *346*, 996–10001. https://doi.org/10.1126/science.1256427.

Fox, A., Carolan, L., Leung, V., Vu, H., Phuong, M., Khvorov, A., Auladell, M., Tseng, Y.-Y., Thai, P.Q., Barr, I., et al. (2022). Opposing Effects of Prior Infection versus Prior Vaccination on Vaccine Immunogenicity against Influenza A(H3N2) Viruses. Viruses *14*, 470. https://doi.org/10.3390/V14030470.

Georgiev, I.S., Doria-Rose, N.A., Zhou, T., Kwon, Y. Do, Staupe, R.P., Moquin, S., Chuang, G.Y., Louder, M.K., Schmidt, S.D., Altae-Tran, 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.

Gouma, S., Kim, K., Weirick, M.E., Gumina, M.E., Branche, A., Topham, D.J., Martin, E.T., Monto, A.S., Cobey, S., and Hensley, S.E. (2020). Middle-aged individuals may be in a perpetual state of H3N2 influenza virus susceptibility. Nat. Commun. *11*, 1–8. https://doi.org/10.1038/s41467-020-18465-x.

Henry, C., Zheng, N.-Y., Huang, M., Cabanov, A., Rojas, K.T., Kaur, K., Andrews, S.F., Palm, A.-K.E., Chen, Y.-Q., Li, Y., et al. (2019). Influenza Virus Vaccination Elicits Poorly Adapted B Cell Responses in Elderly Individuals. Cell Host Microbe *25*, 357-366.e6. https://doi.org/10.1016/J.CHOM.2019.01.002.

Hines, K.E., Middendorf, T.R., and Aldrich, R.W. (2014). Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. J. Gen. Physiol. *143*, 401–416. https://doi.org/10.1085/jgp.201311116.

Keshavan, R.H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. IEEE Trans. Inf. Theory *56*, 2980–2998. https://doi.org/10.1109/TIT.2010.2046205.

Kim, J.H., Davis, W.G., Sambhara, S., and Jacob, J. (2012). Strategies to Alleviate Original Antigenic Sin Responses to Influenza Viruses. Proc. Natl. Acad. Sci. *109*, 13751–13756. https://doi.org/10.1073/pnas.0912458109.

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., Paparoditis, P., Horton, A.P., Frühwirth, A., McDaniel, J.R., Jung, J., Boutz, D.R., Hussein, D.A., Tanno, Y., Pappas, 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.CHOM.2019.01.010.

Lee, J.M., Eguia, R., Zost, S.J., Choudhary, S., Wilson, P.C., Bedford, T., Stevens-Ayers, T., Boeckh, M., Hurt, 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.

Lewnard, J., and Cobey, S. (2018). Immune History and Influenza Vaccine Effectiveness. Vaccines *6*, 28. https://doi.org/10.3390/vaccines6020028.

Marchi, J., Lässig, M., Walczak, A.M., and Mora, T. (2021). Antigenic waves of virus-immune coevolution. Proc. Natl. Acad. Sci. *118*. https://doi.org/10.1073/PNAS.2103398118.

Morris, D.H., Gostic, K.M., Pompei, S., Bedford, T., Łuksza, M., Neher, R.A., Grenfell, B.T., Lässig, M., and McCauley, J.W. (2018). Predictive

Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. Trends Microbiol. *26*, 102–118. https://doi.org/10.1016/J.TIM.2017.09.004.
Ndifon, W. (2011). New methods for analyzing serological data with applications to influenza surveillance. Influenza Other Respi. Viruses *5*, 206–212. https://doi.org/10.1111/j.1750-2659.2010.00192.x.
Petrova, V.N., and Russell, C.A. (2017). The evolution of seasonal influenza viruses. Nat. Rev. Microbiol. *16*, 47–60. https://doi.org/10.1038/nrmicro.2017.118.
Sheng, J., and Wang, S. (2021). Coevolutionary transitions emerging from flexible molecular recognition and eco-evolutionary feedback. IScience *24*, 102861. https://doi.org/10.1016/J.ISCI.2021.102861.
Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, 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.
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. Le, et al. (2021). Age-seroprevalence curves for the multi-strain structure of influenza A virus. Nat. Commun. *12*, 1–9. https://doi.org/10.1038/s41467-021-26948-8.
Zacour, M., Ward, B.J., Brewer, A., Tang, P., Boivin, G., Li, Y., Warhuus, M., McNeil, S.A., LeBlanc, J.J., and Hatchette, T.F. (2016). Standardization of Hemagglutination Inhibition Assay for Influenza Serology Allows for High Reproducibility between Laboratories. Clin. Vaccine Immunol. *23*, 236–242. https://doi.org/10.1128/CVI.00613-15.

SI Figures

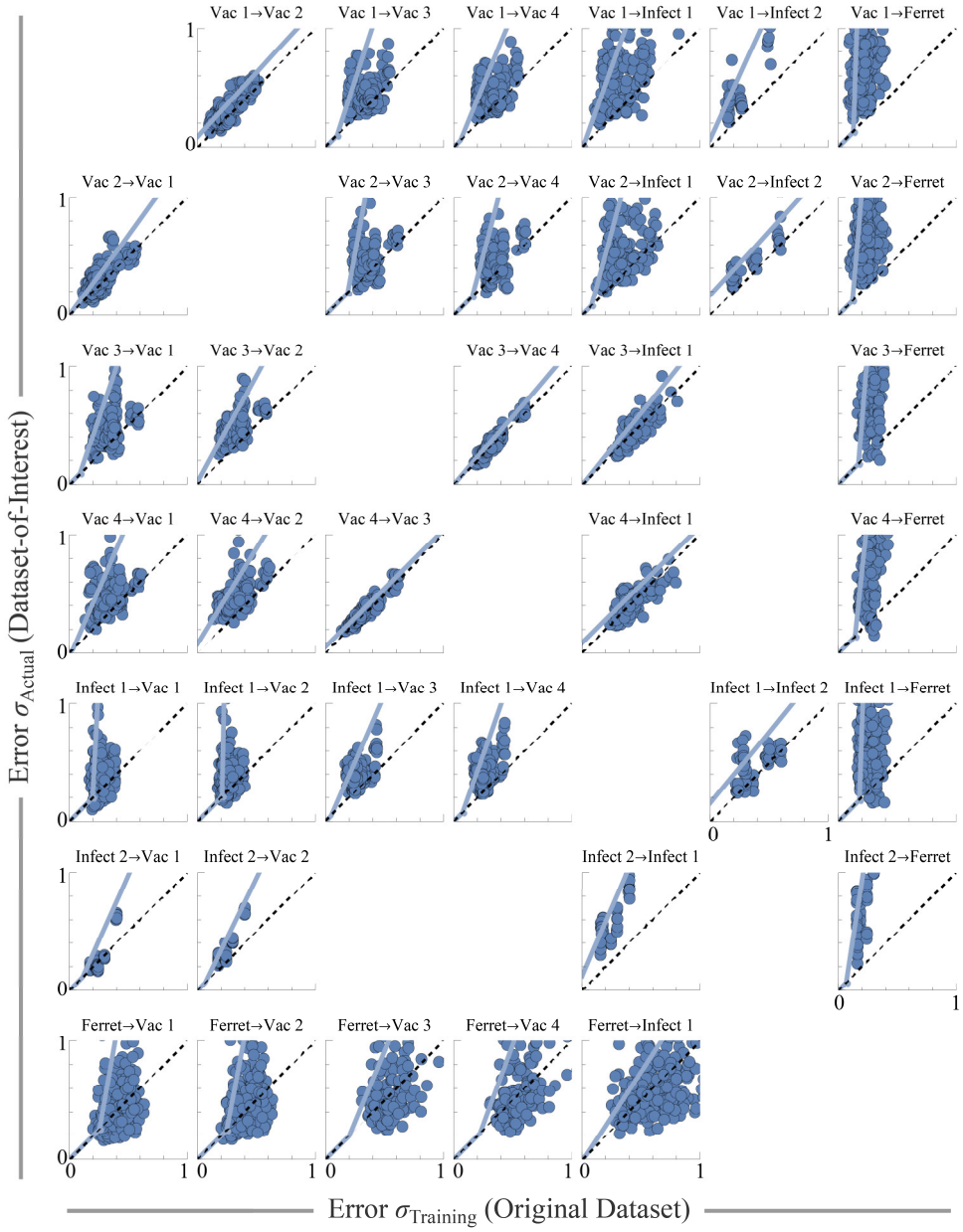

Figure S1. Transferability of behavior between studies examined in this work. Each plot quantifies the transferability relation $f_{j\to k}$ of virus behavior between Dataset j and Dataset k. Each point represents a decision tree that predicts values for a virus V_0 using values from V_1-V_5. Each tree was trained on 30% of samples in Dataset j, with its cross-validation RMSE σ_{Training} computed against the remaining 70% of samples [x-axis]. This tree then predicted this same V_0 using V_1-V_5 in Dataset k, with RMSE σ_{Actual}. Every possible virus (measured in both Dataset j and Dataset k) was withheld and predicted in this manner, and the points shown represent the 5 decision trees with the lowest σ_{Training} (or the top 10 trees if there are fewer than 300 points in the plot, to better capture the relation between datasets). The best-fit perpendicular line $f_\perp$ was fit to the resulting points, and to account for variability (and to overestimate rather than underestimate error) we add to this line the constant f_{RMSE} (the RMSE of the vertical deviations between $f_\perp$ and each point). Lastly, because error should increase when extrapolating the predictions to a new dataset ($\sigma_{\text{Training}} \leq \sigma_{\text{Actual}}$), and because some of the lines are nearly vertical, we enforce that $f_{j\to k}$ lies above $y=x$. In summary, using each plot we define $f_{j\to k}=\max(f_\perp+f_{\text{RMSE}},0)$.

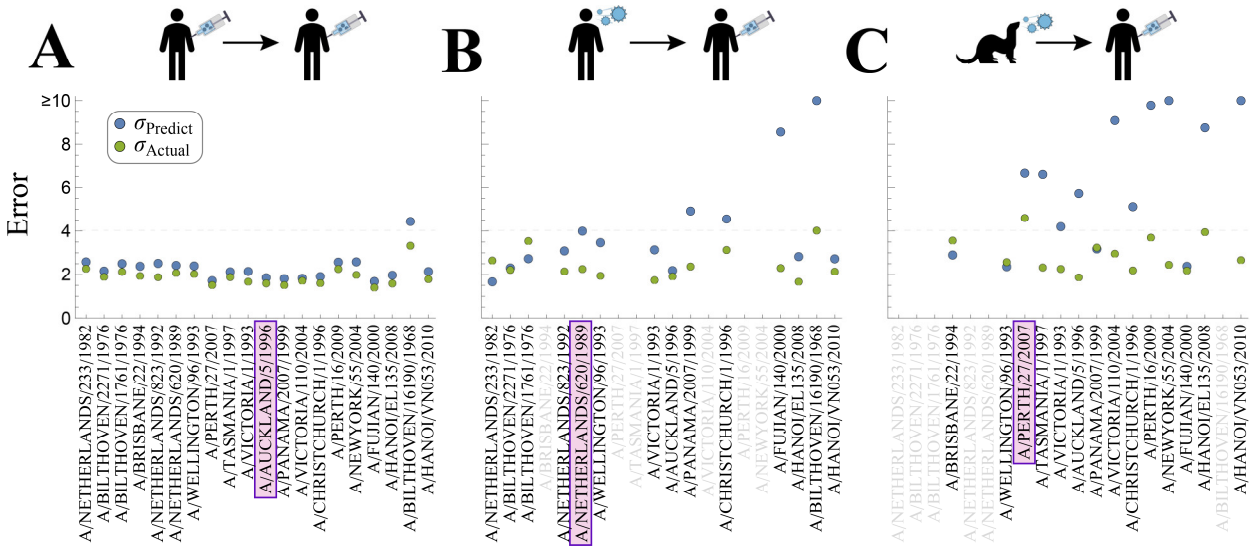

Figure S2. Predicting each virus in Dataset$_{Vac,4}$ using one other dataset. We withhold one virus in Dataset$_{Vac,4}$ (*x*-axis) and predict it using (A) the human vaccination study [Dataset$_{Vac,3}$], (B) the human infection study [Dataset$_{Infect,1}$], or (C) the ferret infection study (Dataset$_{Ferret}$). In each case, we show the estimated error ($\sigma_{Predict}$, blue) and the true error (σ_{Actual}, green). Viruses appear in the same order in each plot, sorted from least-to-greatest average $\sigma_{Predict}$ across the three plots. Grayed-out viruses could not be predicted either because they were absent from a dataset (*e.g.*, Dataset$_{Infect,1}$ did not contain A/BRISBANE/22/1994) or because there was insufficient data. The three viruses shown in Figure 2 are boxed in purple.

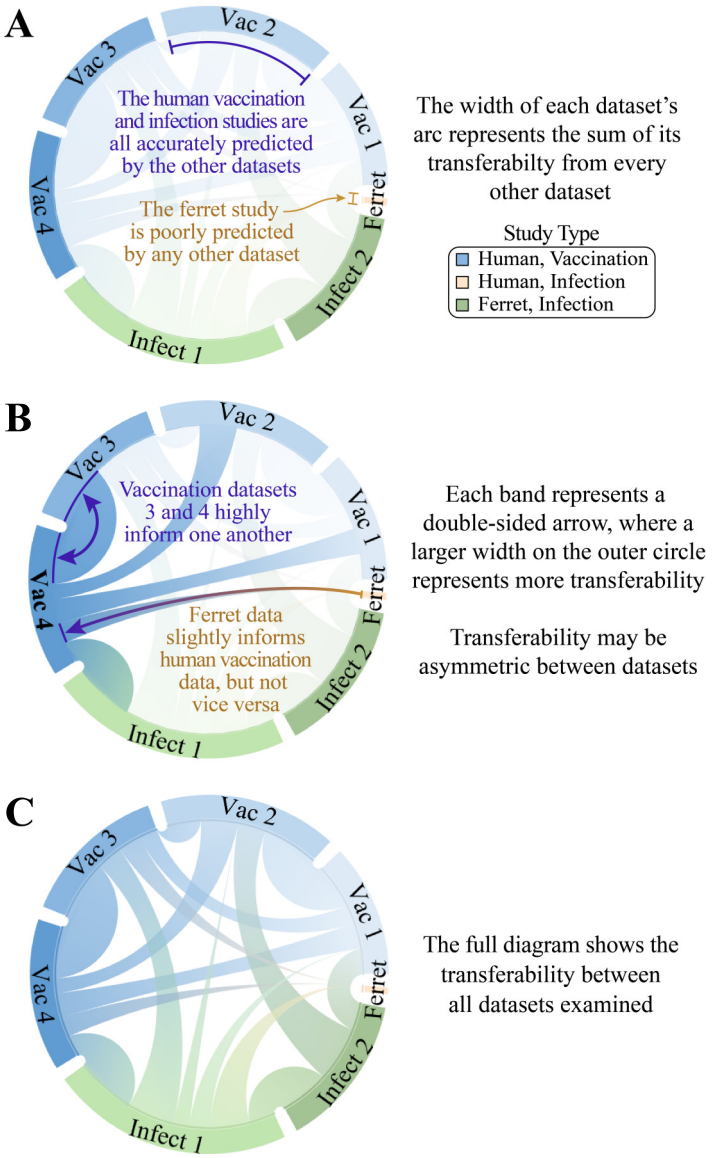

Figure S3. Breaking down the chord diagram depictions of transferability. (A) The chord diagram from Figure 3B representing the transferability between the influenza datasets when considering all data. (B) A wider arc from Study *X*→Study *Y* represents greater transferability. More precisely, transferability equals 1/slope of the linear map in Figure S1, so that studies with near-perfect transferability (slope≈1) will have large width while studies with poor transferability (slope≫1) will have small width. (C) The full diagram from Figure 3B. For cleanliness, we impose a hard threshold and only show transferability when slope≤5 (since greater slopes imply a nearly vertical relationship that represents very poor transferability).

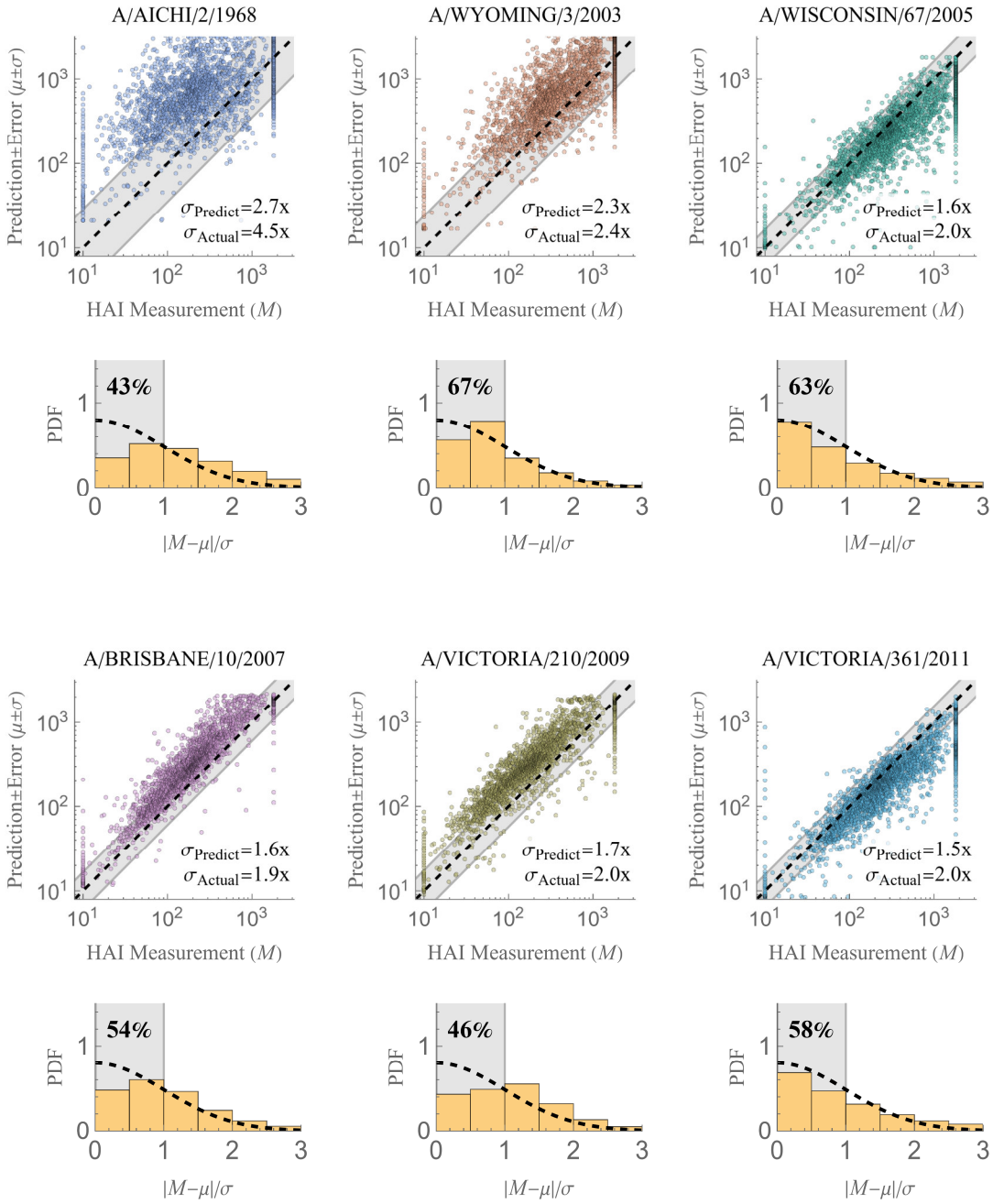

Figure S4. Individual matrix completions in the Vinh dataset. Each of the six Vinh viruses were withheld and predicted using the Fonville datasets. *Scatterplots* show predictions versus measurements. For each virus, the uncertainty of its predictions will be the same for all 25,000 values, and this uncertainty is visualized using the gray bands (showing the fold-error σ_{Predict}); the predicted and true errors are also written at the bottom-right of each plot. For clarity, we only show every 10th data point of the 25,000 measurements, but all statistics are computed using the full data. *Histograms* portray the error distribution for the predictions, with the value in the gray region showing the number of predictions within 1σ of the measurement.

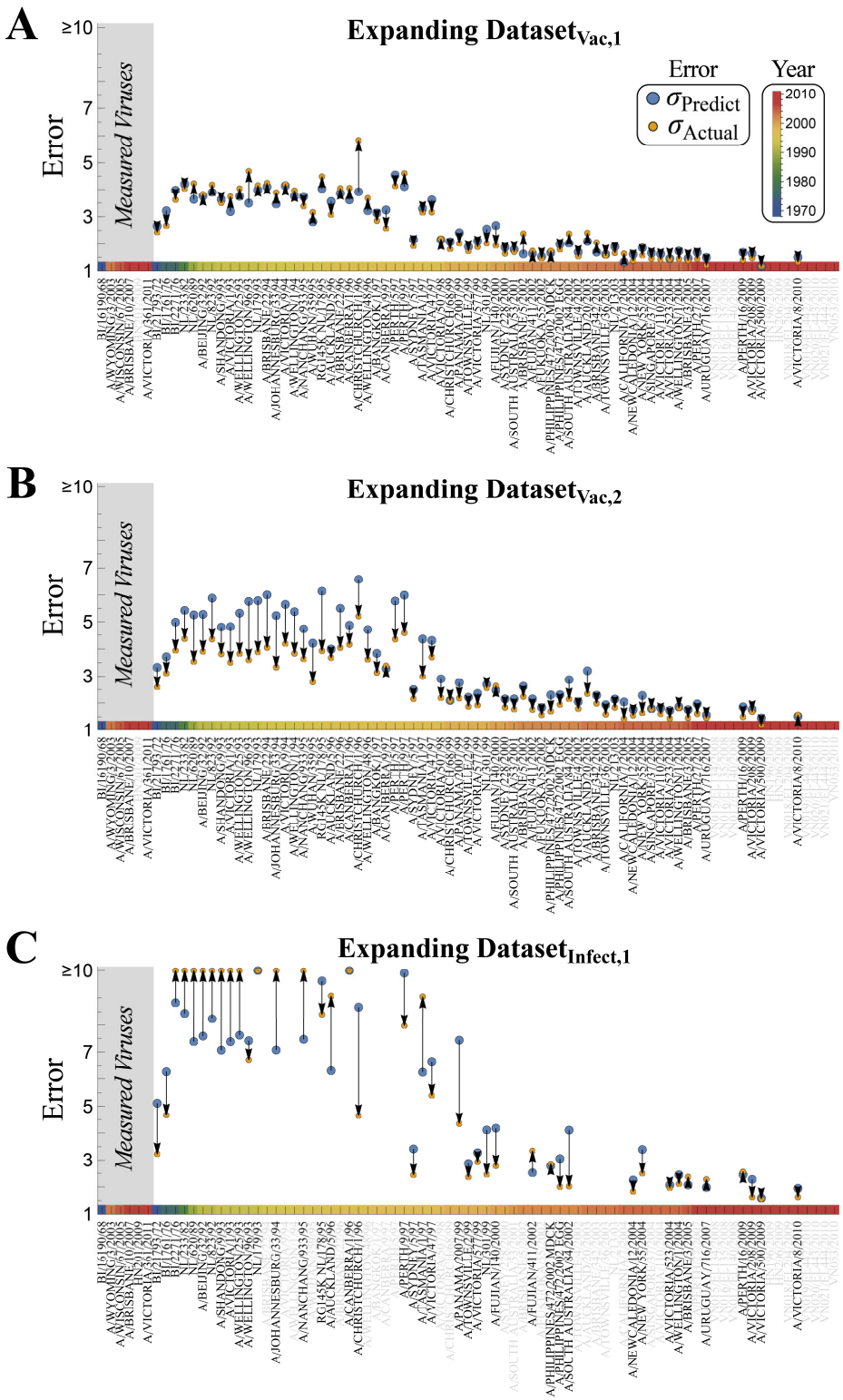

Figure S5. Extrapolating virus behavior in the Fonville datasets using 6 viruses. Analogous to Figure 5, we only use values from the six Vinh viruses (or whatever viruses are present in the Fonville datasets) to predict the behavior of all other viruses. We consider predictions in (A) Dataset$_{Vac,1}$, (B) Dataset$_{Vac,2}$, or (C) Dataset$_{Infect,1}$, which are the three datasets that contribute the most of the Vinh predictions [Figure 3B]. Each plot shows the predicted error [$\sigma_{Predict}$, blue] and true error [σ_{Actual}, gold] of the predictions, with a connecting arrow. Viruses in gray could not be predicted because they either were not in one of the Fonville

Datasets or there was insufficient data to predict them. Viruses from the 1980s and 1990s (which are the furthest away from the 5-6 measured viruses) have the largest error, and this error is slightly overestimated in $\text{Dataset}_{\text{Vac},2}$ and underestimated in $\text{Dataset}_{\text{Infect},1}$. As explained in the Methods, our framework is constructed so that low σ_{Predict} always implies a low σ_{Actual} (with $\sigma_{\text{Predict}} \approx \sigma_{\text{Actual}}$), whereas large σ_{Predict} implies less certainty in σ_{Actual}. A good rule of thumb from these results is to not use values with a predicted error ≥6-fold, since their true error may be even larger; we note that all new values in Figure 5 have a predicted error <6-fold.

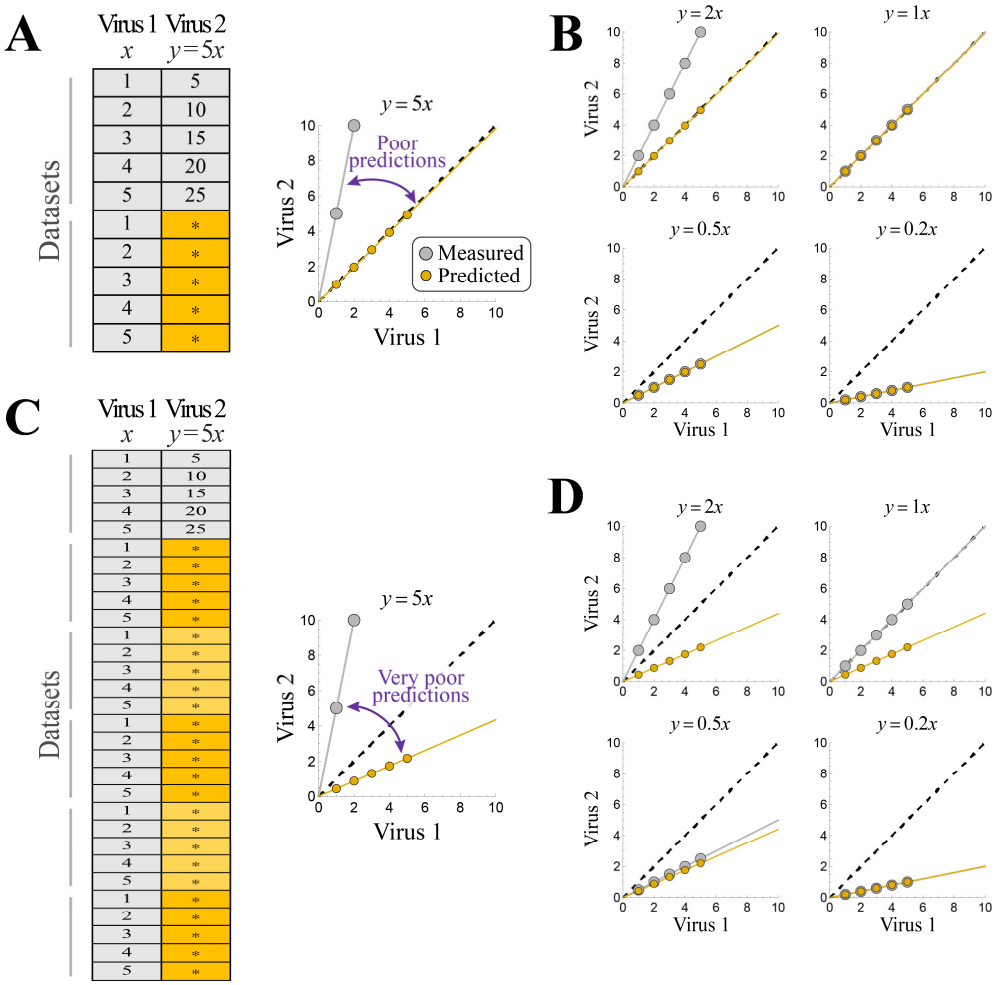

Figure S6. Nuclear norm minimization can fail in a simple, noise-free setting. (A) A noise-free toy example where the measurements of two viruses are proportional to each other (y=5x), but Virus 2 is incorrectly predicted as y=x. (B) This problem holds for any relation y=mx where m>1, although values of m≤1 lead to perfect recovery. (C,D) The same setup with n copies of the missing measurements. The problem is now exacerbated, with Virus 2 predicted as $y=n^{-1/2}x$ whenever $m>n^{-1/2}$.

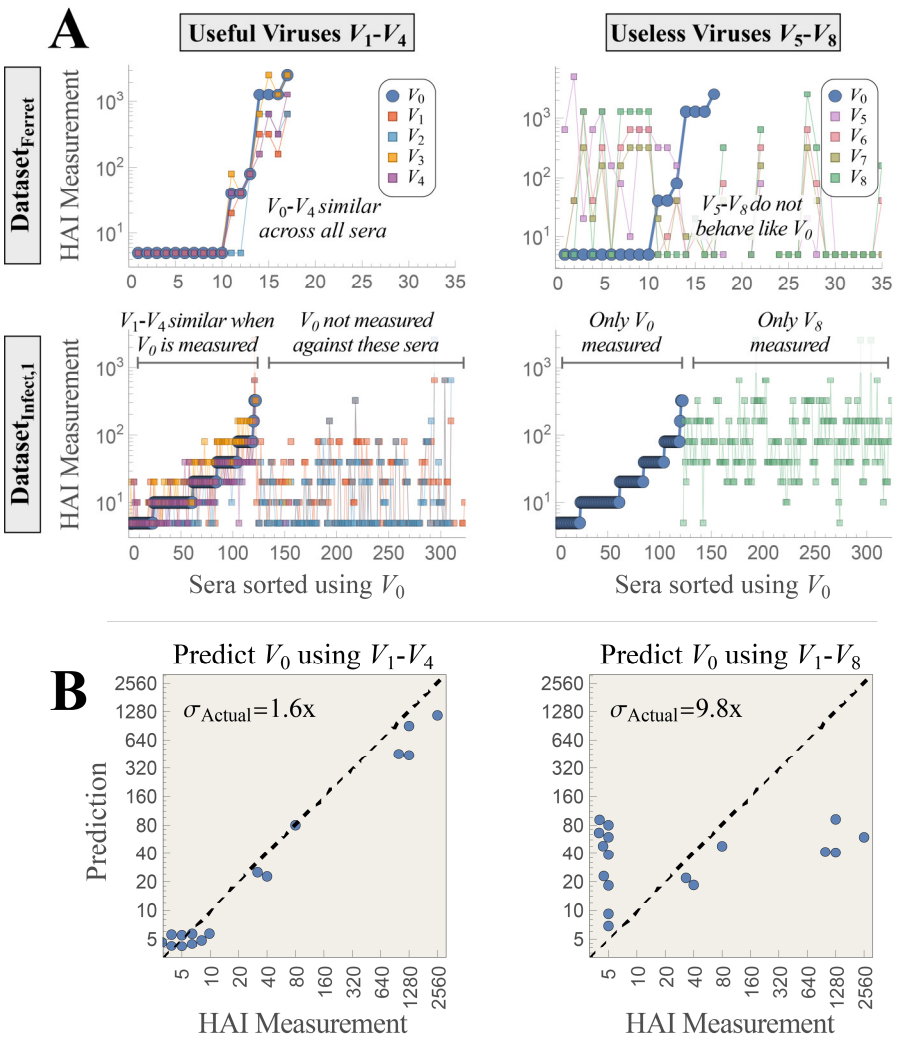

Figure S7. Nuclear norm minimization may give poor predictions when there are large swaths of missing values. Predictions for virus V_0 [specified below] from Datasets$_{Infect,1 \rightarrow Ferret}$ are highly accurate when using the "useful" viruses V_1-V_4 that behave similarly in both studies, but highly inaccurate when adding the additional "useless" viruses V_5-V_8 that don't behave like V_0 in either study. (A) Plot of the titers of the useful and useless viruses in both datasets, with sera sorted according to the HAI titers of V_0. Values for V_1-V_4 closely match those of V_0 for all sera in Dataset$_{Ferret}$ and for all sera where V_0 is measured in Dataset$_{Infect,1}$ (the first 125 sera). In contrast, V_5-V_8 do not behave like V_0 in Dataset$_{Ferret}$; in Dataset$_{Infect,1}$ viruses V_5-V_7 are never measured, and V_8 is only measured against sera where V_0 was not measured. Hence, V_5-V_8 should ideally not influence the matrix completion of V_0. (B) The resulting predictions vs measurements for V_0 only using V_1-V_4 [left] or using both V_1-V_4 and V_5-V_8 [right], with the latter leading to significantly larger error. In the Fonville datasets, these viruses represent V_0=VN018/EL204/2009, V_1-V_4={HN201/2009, HN206/2009, VN019/EL442/2010, VN020/EL443/2010}, V_5-V_8={A/SINGAPORE/37/2004, A/SOUTH AUSTRALIA/53/2001, A/SYDNEY/228/2000, A/SOUTH AUSTRALIA/84/2002}.