Reviewed Preprint
v1 • February 26, 2025
Not revised

Reviewed Preprint
v2 • July 24, 2026
Revised by authors

# Permute-match tests detect significant correlations between time series despite nonstationarity and limited replicates

**Alex E Yuan**[1], **Wenying Shou**[2]

[1]Molecular and Cellular Biology PhD program, University of Washington and Fred Hutchinson Cancer Center, Seattle, United States • [2]Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

**For correspondence:** alexericyuan@gmail.com

**Competing interests:** No competing interests declared

**Funding:** See page 30

**Reviewing editor:** Sandeep Krishna, National Centre for Biological Sciences-Tata Institute of Fundamental Research, India

© 2025, Yuan & Shou. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

## eLife Assessment

This manuscript reports an **important** new statistical method for calculating the significance of correlations between two time-series, which provides more accuracy than other methods when the data has few replicates. The proposed method solves a real-life problem that is frequently encountered and is broadly applicable to many realistic datasets in many experimental contexts. The technique is supported with **compelling** mathematical derivations as well as analysis of both computer-generated and previously published experimental data.

https://doi.org/10.7554/eLife.103703.2.sa3

## Abstract

Researchers frequently analyze correlations between pairs of time series by determining whether an observed correlation is stronger than expected under the null hypothesis of independence. However, the time series are often nonstationary, with statistical properties that change over time, thereby making standard tests invalid. If sufficient replicates exist, a trial-swapping permutation test can be performed that handles nonstationarity by comparing within-replicate correlations to between-replicate correlations. Although largely assumption-free, this test is fundamentally limited by the number of replicates ($n$) because its minimum p-value is $1/n!$. With $n = 3$, this minimum is 1/6, rendering thresholds like 0.05 unattainable. This limits its use considerably in animal experiments, where $n$ may be as low as 3. We propose permute-match tests — modified permutation tests that can report lower p-values of $2/n^n$ or $1/n^n$ under strong evidence of dependence. Permute-match tests guarantee a false positive rate at or below the significance level when replicates are independent and identically distributed. The bound of $1/n^n$ is not gratuitously conservative, since it cannot be further lowered without additional assumptions. We demonstrate our approach using synthetic data and apply it to an existing dataset with 3 independent groups of zebrafish, confirming the observation that zebrafish swim faster when directionally aligned.

## Introduction

Scientists frequently look for correlations between variables to identify potentially important relationships or to support conceptual models. In disciplines with a focus on dynamics such as ecology and physiology, it is common to measure correlations between time series. Many important biological processes are nonstationary, meaning that their statistical properties (e.g., mean or variance) change systematically across time. Such processes range from the spread of an invasive species to a cell's response to a new environmental stress.

Interpreting a correlation between a pair of nonstationary time series can be highly fraught because it is easy to obtain a seemingly impressive correlation between two time series that have no meaningful relationship [1]. For example, the sizes of any two exponentially-growing populations will be correlated over time due to a shared growth law, even if the populations lack any interactions or shared influences (e.g., bacterial cultures in separate tubes). To avoid overinterpreting spurious correlations, it helps to distinguish between the concepts of "correlation" and "dependence". In time series research, the term "correlation" is often used procedurally [2, 3, 4]. Similarly, here we define a correlation function ($\rho$) to be any function that takes two time series and produces a statistic, although it is usually interpreted as a measure of similarity or relatedness.

In contrast to correlation, "dependence" is a hypothesis about the relationship between variables, and it has clearer scientific implications. Two events $A$ and $B$ are called dependent if $P(A, B)$, the probability that they both occur, differs from $P(A)P(B)$, the product of their individual probabilities. The formal definition of dependence extends this idea to continuous variables, where discrete probabilities are replaced by probability densities [5, 6]. Similarly, two temporal processes are dependent if the probability of observing any particular pair of trajectories differs from the product of the probabilities of the individual trajectories. The importance of dependence relationships can be attributed to Reichenbach's common cause principle, which states that if two variables are dependent, then either they share a common causal influence, or one variable causally influences the other (possibly indirectly) [7, 8]. Thus, it is useful to first test whether an observed correlation indicates dependence before pursuing any mechanistic explanations.

There are a handful of systematic approaches to testing for dependence between nonstationary time series. However, most are fairly limited in their scope. First, when possible, a nonstationary time series can be transformed to become stationary, meaning that its statistical properties do not change systematically across time. This enables access to a wide arsenal of tests applicable to stationary data. Transformations include subtracting a trend, taking the derivative (more precisely, "differencing" between neighboring points), or choosing a stationary-seeming window of time [9, 10]. However, it is easy to see potential pathologies in each of these: Taking the derivative of an exponential curve just produces another exponential curve; subtracting a fitted linear trend from the random walk $X(t) = X(t-1) + \epsilon(t)$ (where $\epsilon(t)$ is random noise) does not make it stationary [9]; similarly, searching for a stationary window of the same random walk process is futile since the variance of $X(t)$ increases with each step. In a second approach, one may compare the observed correlation with correlations obtained when one time series is replaced by a synthetic replicate, where synthetic replicates are generated in a way that reflects the original nonstationary process [11]. However, these require a correct model of how the statistical properties of the process evolve over time. Lastly, there are tests within the econometrics literature that can provide evidence for dependence between random-walk-like nonstationary processes by detecting a property called cointegration, but cointegration only occurs when some linear combination of the time series is stationary [12].

Alternatively, a permutation-based approach is possible when there are multiple identically distributed and independent (iid) replicates (or trials; we use "replicates" and "trials" interchangeably). In this case, one may evaluate the significance of a within-replicate correlation by comparing it to between-replicate correlations. That is, if $X_i$ and $Y_i$ are time series of variables $X$ and $Y$ from replicate $i$, the correlation of $(X_i, Y_i)$ may be compared to the correlation of $(X_i, Y_{j\neq i})$. If the two variables are dependent, within-replicate correlations may differ systematically from between-replicate correlations. This approach is sometimes called "inter-subject surrogates" [13, 14].

The inter-subject surrogate approach can be used to test for dependence in each trial separately [15]. In this case, a simple nonparametric test of dependence between, for instance, $X_1$ and $Y_1$ can be performed by computing the correlation $\rho(X_1, Y_j)$ for $j = 1, \ldots, n$ and writing down a $p$-value as the proportion of these correlations that are greater than or equal to $\rho(X_1, Y_1)$ [16, 17]. For this

approach, at least $n$ = 20 trials are needed if one wishes to possibly obtain a $p$-value of 0.05 or below. This procedure checks for dependence in trial 1 specifically; assessing dependence in a different trial via this method would involve an analogous test.

To reduce the required number of trials below 20, a straightforward strategy is to perform a single permutation test (Fig 1A) using the data from all replicates [18, 19]. As an example with $n$ = 4 trials, the test procedure begins by computing the mean within-trial correlation:

$$\rho_{\text{orig}} = \frac{1}{4}\left(\rho\left(X_1, Y_1\right) + \rho\left(X_2, Y_2\right) + \rho\left(X_3, Y_3\right) + \rho\left(X_4, Y_4\right)\right).$$

Next, as a null model, recompute the mean correlation after permuting the $Y$ time series while holding the $X$ time series in the original order. For instance, one such permutation might be:

$$\rho_{\text{shuffle}} = \frac{1}{4}\left(\rho\left(X_1, Y_3\right) + \rho\left(X_2, Y_2\right) + \rho\left(X_3, Y_4\right) + \rho\left(X_4, Y_1\right)\right).$$

A $p$-value can be calculated as the proportion of permutations (including the original ordering) that produce a mean correlation that is as strong as, or stronger than, $\rho_{\text{orig}}$ (Fig 1A). One may then detect a correlation at the significance level $\alpha$ if $p \leq \alpha$. We emphasize that these permutations are obtained by swapping trials, not by swapping time points. This test has been used to detect correlations between time series in neuroscience and psychology settings [20, 18, 21]. It has also been used to detect correlations between variables measured in brain images, which can have similar nonstationarity challenges as time series [22]. A noteworthy advantage of this approach is that it is valid (meaning that the false positive rate is guaranteed to not exceed $\alpha$) under very mild assumptions. Namely, the test is valid if the $X_i$ trials are exchangeable with one another (i.e. all permutations of the sequence $X_1$, ..., $X_n$ have the same joint probability distribution), or if the $Y_i$ trials are exchangeable (see Corollary 4 in Appendix 1).

Yet, even the permutation test remains considerably limited by the number of trials. The lowest $p$-value that can be obtained with $n$ replicates is $1/n!$ since there are $n!$ possible permutations. For example, with $n$ = 3, the lowest possible $p$-value is 1/3! ≈ 0.17 and with $n$ = 4, the lowest is 1/4! ≈ 0.04. These minima apply even if the average within-trial correlation is much greater than the average between-trial correlation. While replicate counts of 4 and 5 can produce $p$-values below the conventional 0.05 threshold, these $p$-values remain modest. This limitation is particularly problematic when multiple hypothesis testing is required, as the modest $p$-values may no longer be significant after applying corrections such as the Bonferroni method.

The challenge of limited replicates is a reality in many experimental settings due to cost and feasibility concerns [23, 24, 25]. Factors outside the control of investigators further restrict the number of useful replicates. For example, a study of cerebral ischemia in pigs began with 6 animals, but lost 2 due to death and experimental mishap, so complete records exist for only 4 pigs [26]. Similarly, in an animal migration study, authors affixed GPS tags to 13 juvenile cuckoos, but only 5 of them produced GPS records suitable for the study due to heterogeneous behavior among cuckoos [27]. Finally, secondary analysis of public data from earlier studies can be a resource-efficient mode of research [28], although this approach is necessarily limited to the number of replicates within the existing dataset.

In this article, we show that permute-match tests, by adding additional steps to the permutation test, can achieve $p$-values as low as $1/n^n$. For $n$ = 3 or 4 respectively, the lowest possible $p$-value is then ≈ 0.04 or 0.004, compared to 1/3! ≈ 0.17 or 1/4! ≈ 0.04. The permute-match tests only require that the replicates are iid. Thus, permute-match tests allow the data analyst to detect dependence with stronger confidence when there is sufficient evidence to do so. We illustrate this approach using simulations as well as a real world example involving recordings of zebrafish swimming behavior with only 3 replicates.

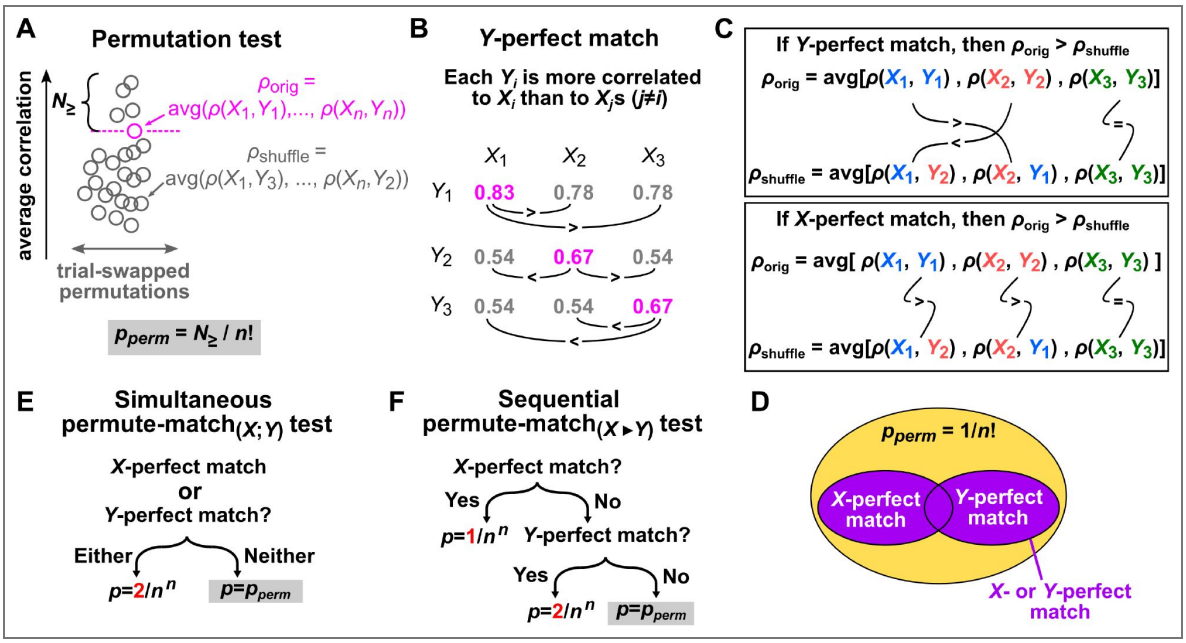


**Figure 1. The permutation test, perfect match concept, and permute-match tests.**

(**A**) The permutation test. Circles indicate the average correlation without trial swapping ($\rho_{orig}$; pink) or with trial swapping ($\rho_{shuffle}$; grey). Here, $N_{\geq}$ is the number of average correlations (including $\rho_{orig}$) that are equal to or larger than the original correlation $\rho_{orig}$. It follows that $p_{perm} = N_{\geq}/n!$ because $n!$ is the number of ways the trials can be shuffled. (**B**) Illustration of a *Y*-perfect match, represented as a table of correlations wherein each diagonal entry is the greatest in its own row. Each line with a ">" or "<" symbol denotes a required ordering relationship between the two numbers at either end of the line. Note that this example has a *Y*-perfect match, but not an *X*-perfect match. For an *X*-perfect match, each diagonal entry would need to be the greatest in its own column. (**C**) If an *X*- or *Y*-perfect match occurs, then the original correlation $\rho_{orig}$ is always greater than any shuffled correlation $\rho_{shuffle}$. Top: If a *Y*-perfect match occurs, then each $Y_i$ gives the greatest correlation when it is paired with the $X_i$ from the same trial. A shuffled correlation $\rho_{shuffle}$ pairs $Y_i$s from some trials with $X_j$s from different trials, thereby reducing the correlation. Lines with symbols ">", "<", and "=" denote comparisons between terms. Bottom: If an *X*-perfect match occurs, a similar argument can be applied. (D) Since a perfect match ensures that the original correlation is greater than any shuffled correlation, a perfect match also ensures that $p_{perm}$ takes on its lowest possible value of $1/n!$. (E) The simultaneous permute-match$_{(X;Y)}$ test. (F) The sequential permute-match$_{(X \blacktriangleright Y)}$ test; note that the permute-match$_{(Y \blacktriangleright X)}$ test is defined analogously.

# Results

## The concept of an *X*-perfect or *Y*-perfect match

Central to our approach is the concept of an *X*-perfect match or *Y*-perfect match. To motivate the ideas, suppose that a group of $n$ graduate students $\boldsymbol{X} = (X_1, \ldots, X_n)$ has been paired with a group of $n$ advisors $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ so that the $i$th student ($X_i$) is paired with the $i$th advisor ($Y_i$). Moreover, let $\rho(X_i, Y_j)$ be a number that measures how well the $i$th student and the $j$th advisor get along. We say that the $i$th student is "happy" if they get along with their own advisor strictly better than they get along with any other advisor, meaning that $\rho(X_i, Y_i) > \rho(X_i, Y_j)$ for all $j \neq i$. Similarly, the $i$th advisor is "happy" if $\rho(X_i, Y_i) > \rho(X_j, Y_i)$ for all $j \neq i$. The arrangement of student-advisor pairs is *X*-perfect if all students are happy, and *Y*-perfect if all advisors are happy.

Analogously, if $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are collections of time series with $n$ trials each, and if $\rho$ is some correlation function, we say that an *X*-perfect match has occurred if (and only if) for each $i \in \{1, \ldots, n\}$,

$$\rho\left(X_i, Y_i\right) > \rho\left(X_i, Y_j\right) \text{ for all } j \neq i.$$

Similarly, a *Y*-perfect match has occurred if and only if for each $i$,

$$\rho\left(X_i, Y_i\right) > \rho\left(X_j, Y_i\right) \text{ for all } j \neq i.$$

See Fig 1B for an example of a *Y*-perfect (but not *X*-perfect) match.

Throughout the article, we use boldface $\boldsymbol{X}$ or $\boldsymbol{Y}$ to indicate a collection of trials, $X_i$ or $Y_i$ to indicate an individual trial, and $X$ or $Y$ to refer to the variable in general.

A perfect match (either *X*-perfect or *Y*-perfect) provides especially strong evidence to support the hypothesis that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent. With a perfect match, the original correlation $\rho_{\text{orig}}$ is strictly greater than any of the shuffled correlations, as illustrated in Fig 1C. Thus, a perfect match guarantees $p_{perm} = 1/n!$, the lowest $p$-value achievable with the permutation test (Fig 1D). Furthermore, letting $P$ denote probability, we can prove that

$$P(X\text{-perfect match}) \leq 1/n^n$$

under the null hypothesis $H_0$ wherein (1) the $Y_i$ trials are iid, (2) the $X_i$ trials are iid, and (3) $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent (Lemma 8 in Appendix 2). In Proposition 14 (Appendix 2), we show that this upper bound of $1/n^n$ cannot be reduced further without requiring additional assumptions.

The same bound applies to the *Y*-perfect match: under $H_0$ we have $P$ ($Y$ −perfect match) ≤ $1/n^n$ (Lemma 6 in Appendix 2) and thus

$$P(X\text{-perfect match or } Y\text{-perfect match}) \leq 2/n^n. \quad (1)$$

## Permute-match tests: modified permutation tests of dependence

We now describe tests for dependence between nonstationary time series based on the ideas just laid out, which we will call permute-match tests. Suppose that an experiment with $n$ replicates produces time series $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, and we want to test whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent. For example, we may have video recordings of $n$ animals, each in a separate but identically constructed enclosure; for the $i$th animal, we extract a time series of movement speed ($X_i$) and a time series of the distance from a light source ($Y_i$). Using these data, we may wish to investigate whether the animals under study tend to move at different speeds depending on their

distance from the light source. We introduce two tests: simultaneous permute-match and sequential permute-match. As we will see, the former can be more reproducible while the latter can be more powerful.

In the simultaneous permute-match test (Fig 1E), we check for both an $X$-perfect and a $Y$-perfect match. If either the $X$- or $Y$-perfect match occurs, we write down a $p$-value to be $p = 2/n^n$ as per equation 1. If neither $X$ nor $Y$ achieves a perfect match, then we fall back to the permutation test: $p = p_{perm} = N_{\geq}/n!$ where $N_{\geq}$ is the number of all $n!$ possible permutations (including the original ordering) that produce an average correlation that is at least as large as the original ordering (Fig 1A).

In the sequential permute-match test, we begin by choosing a variable (either $X$ or $Y$) for the first perfect match test. For example, suppose we choose $X$ first (the "permute-match$_{(X \blacktriangleright Y)}$" test). If an $X$-perfect match is observed, then $p = 1/n^n$ (Fig 1F). If $X$-perfect match is not observed, we then check for a $Y$-perfect match, and write $p = 2/n^n$ if a $Y$-perfect match is observed. If neither an $X$-perfect nor $Y$-perfect match occurs, then $p = p_{perm}$. A permute-match$_{(Y \blacktriangleright X)}$ test is defined analogously.

Under $H_0$ (i.e. the $Y_i$ replicates are iid, the $X_i$ replicates are iid, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent), all the above tests are valid, meaning that for any significance level $\alpha$, we have $P(p \leq \alpha) \leq \alpha$. This fact is proven as Theorems 11-13 in Appendix 2.

The permute-match tests (Fig 1E-F) are "upgrades" to the permutation test (Fig 1A) because permute-match tests always report the same or a lower $p$-value (since $1/n^n < 2/n^n \leq 1/n!$ for all integers $n \geq 2$; see Lemma 10 in Appendix 2). In exchange for lower $p$-values, the permute-match tests have a slightly more restrictive data requirement than the permutation test. The permutation test is valid either if the $X_i$ trials are exchangeable, or if the $Y_i$ trials are exchangeable (Corollary 4 in Appendix 1). The permute-match tests require that all $X_i$ trials are iid and that $Y_i$ trials are iid. If trials are iid, then they are necessarily exchangeable. However, exchangeable trials are not always iid. For example, the first five cards drawn from a shuffled deck are exchangeable because all possible orderings of cards are equally likely, but they are not iid because if the first card is an ace of hearts, the second card cannot be the ace of hearts. In practice, biological experimental designs typically use replicates that are both exchangeable and iid, so the more restrictive requirement of the permute-match test is often inconsequential.

The different tests – permutation test, simultaneous permute-match test, and sequential permute-match tests – are valid tests of dependence given iid replicates (Theorems 11-13 in Appendix 2). In other words, they all guarantee false positive rates at or below the significance level. Two subtle aspects of the permute-match tests are worth mentioning here. First and crucially, the permutation test and perfect match events are *nested*. That is, an $X$- or $Y$-perfect match is possible only if $p_{perm}$ is already at its lowest possible value of $1/n!$ (Fig 1C-D). We can therefore combine the checks for perfect matches with the permutation test without inflating the false positive rate above the nominal level. This would not work if the tests were not nested, because then each test would separately contribute false positive results. The second subtlety refers specifically to the sequential permute-match tests. These tests use the lower $p$-value of $1/n^n$ if the first perfect match occurs, but use the more conservative "Bonferroni-corrected" $p$-value of $2/n^n$ if only the second match occurs (Fig 1F). This superficially resembles the practice of conducting additional tests merely because earlier tests did not result in a detection – a form of $p$-hacking that generally results in an invalid final $p$-value. However, our sequential permute-match procedure is actually valid because it combines binary events instead of continuous variables. See Appendix 3 for a detailed discussion of this point.

## Permutation and permute-match test variants may differ in statistical power

Although the permutation test, simultaneous permute-match test, and sequential permute-match test are all valid, they vary in power – the probability of detecting true dependence. Depending on the number of replicates $n$ and the desired significance level $\alpha$, there are four regimes to consider

(Fig 2A).

Regime 1: when $\alpha \geq 1/n!$, all tests have the same power. If the permutation test detects dependence, permute-match tests will too since they fall back to a permutation test in the "worst case" (Fig 2A, column 1). Conversely, if the permutation test did not detect dependence (i.e. $p_{perm} > \alpha$), then it follows that $p_{perm} > 1/n!$, which implies that neither an $X$-perfect nor $Y$-perfect match occurred (Fig 1D), and so the permute-match tests cannot have detected dependence either.

Regime 2: when $1/n! > \alpha \geq 2/n^n$, all permute-match tests are tied for highest power (Fig 2A, column 2). In this regime, the permutation test has zero power, since its $p$-value is limited at $1/n!$. Conversely, if either an $X$-perfect or a $Y$-perfect match occurs, then, the sequential permute-match tests as well as the simultaneous permute-match$_{(X;Y)}$ test will all detect dependence.

Regime 3: when $2/n^n > \alpha \geq 1/n^n$, the simultaneous permute-match$_{(X;Y)}$ test will have no power as the desired $p$-value is below the lowest possible $p$-value achievable by the test. Only the permute-match$_{(X \blacktriangleright Y)}$ or $_{(Y \blacktriangleright X)}$ tests now have power (Fig 2A, column 3). Note that permute-match$_{(X \blacktriangleright Y)}$ and permute-match$_{(Y \blacktriangleright X)}$ test may differ in power, as can be seen in the upper left panel of Fig 2D. We do not currently know how to pre-diagnose which of these two tests will be most powerful based on a given dataset, so the choice of which test to run in this case is arbitrary.

Regime 4: when $\alpha < 1/n^n$, no test has any power even when a perfect match occurs.

Figure 2B-D illustrates these different scenarios using a simulated nonstationary system and some common significance cutoffs. We simulated time series of two variables ($X$ and $Y$) from a small number of replicates (between 3 and 5). Here, the time series were produced from a pair of linear time trends with coupled additive noise ($r_{X,Y}$ determines the strength of coupling; Fig 2B). We chose the Pearson correlation coefficient as our correlation function $\rho$. The various sub-panels of Fig 2D illustrate the different aforementioned regimes.

The above example was chosen for ease of presentation: in fact, the example is so simple that it may be reasonably treated by just fitting and removing the linear trend. In Appendix 4, we show the overall higher statistical power of permute-match tests over the permutation test in a more complex example where the data-generating process is nonstationary (a logistic map with time-varying parameters) and where a nonlinear correlation function (cross-map skill) is preferable to (linear) Pearson correlation.

## The permute-match test's false positive rate depends on the process tested

The permute-match test is conservative – meaning that its false positive rate can fall below the significance level – because both the permutation test and perfect match test are conservative. As discussed elsewhere [29], the permutation test is conservative when $\alpha$ is not one of its possible $p$-values and when ties may occur between the original correlation and shuffled correlations. Checking for a perfect match is similarly conservative. The actual probability of a false-alarm perfect match event can vary depending on the process being tested.

To see this consider the permute-match$_{(Y \blacktriangleright X)}$ test in the setting where $n = 3$, where $\alpha = 0.05$, and where $r_{X,Y} = 0$ (independent $X$ and $Y$). Note that in this case, obtaining a $Y$-perfect match (and thus $p = 1/n^n$) is necessary and sufficient to detect dependence since $\alpha$ is too low for detection by either the permutation test or the $p = 2/n^n$ leg of the permute-match test. In the linear system of Fig 2, we observed among 5000 simulations a detection rate of 0.0148, significantly below the upper bound of $1/3^3$ (Figure 2 – Source data 1; one-tailed exact binomial test, $p < 10^{-20}$). Conversely, in the nonlinear system of Fig S2, this same event ($Y$-perfect match under $n = 3$ and $r_{X,Y} = 0$) occurs with a detection rate of 0.0328, not significantly below the upper bound of $1/3^3$ (Figure S2 - Source data 1; one-tailed exact binomial test, $p = 0.059$). Thus, depending on the underlying process studied, the actual chance of a perfect match happening under independence may be near or significantly below the theoretical upper bound.

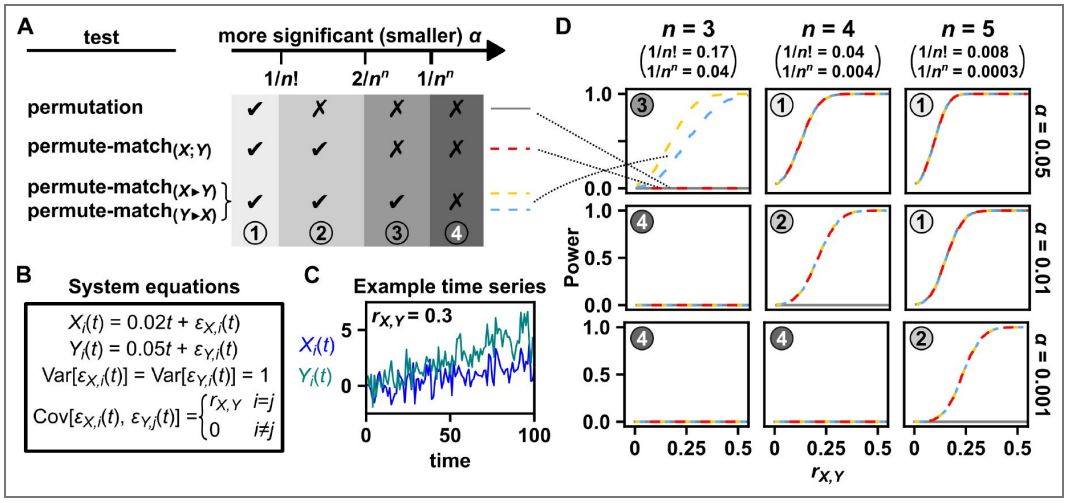


**Figure 2. Statistical power of various permutation test variants as illustrated using a simple nonstationary system.**

(**A**) Summary of test power as a function of the significance level $\alpha$ and the number of replicates $n$. (**B, C**) System equations and example dynamics. The processes $X$ and $Y$ are given by a linear trend with additive noise on a time grid $t = 1, 2, \ldots, 100$. The noise terms $\epsilon_{X,i}(t)$ and $\epsilon_{Y,i}(t)$ are drawn from a bivariate normal distribution with a mean of 0, variance of 1, and covariance of $r_{X,Y}$. The chart shows an example pair of time series where $X$ and $Y$ are dependent ($r_{X,Y} = 0.3$). (**D**) Statistical power for the permutation test and various permute-match tests as a function of the replicate number $n$, significance level $\alpha$, and strength of dependence $r_{X,Y}$. Power was estimated as the proportion of simulations in which dependence was detected, calculated from 5000 simulations at each value of $r_{X,Y}$ between $r_{X,Y} = 0$ and 0.54 in steps of size 0.01. At $r_{X,Y} = 0$, there is no dependence, so the curve at that point indicates the false positive rate rather than power. We chose the Pearson correlation coefficient as our correlation function $\rho$. See (**A**) for the color legend.

## An example from animal behavior science

Living in groups is a common experience among animals. For example, at least half of fish species are thought to form groups at some stage of life [30, 31]. The properties of these groups can impart a variety of important fitness effects [32]. In groups of fish, coordinated swimming may facilitate foraging (e.g., by enabling fish to "communicate" the location of food to others) and predator escape (e.g., by confusing predators) [33, 31]. One study [31] observed, in videos of small groups of zebrafish (8 fish per group), that the fish swam faster when in a high-polarization state (i.e. directionally aligned) than when in a low-polarization state. A correlation between polarization and speed might indicate statistical dependence between the two variables, or it might just be a consequence of temporal trends (e.g., polarization and swimming speed both happen to decrease with time).

Using a publicly available dataset [34] containing fish trajectories from 3 replicate 10-minute videos of small groups of juvenile *Danio rerio* zebrafish (10 fish per group), we performed a permute-match test of dependence between polarization and fish swimming speed. As we will see, although swimming speed and the extent of polarization are potentially nonstationary, the permute-match test nevertheless detects a statistical dependence between them.

Directional polarization is quantified as the mean resultant length $C_t$ of fish velocity [35]. We represent the swimming direction of each fish by a unit vector, and then define $C_t$ as the length of the average unit vector (Fig. 4; Eq. 2). The mean resultant length is bounded between 0 (e.g., two opposite vectors) and 1 (e.g., two parallel vectors). To quantify speed, we took the "average individual speed", which is a term we use to denote an average across individuals, not across time. More precisely, the average individual speed of a 10-fish group at time $t$ is $(s_{1,t} + s_{2,t} + \cdots + s_{10,t})/10$ where $s_{i,t}$ is the speed of fish $i$ at time $t$. Note that average individual speed is distinct from the speed of the group center, which has also been studied in *Danio* fish [32].

Neither the average individual speed nor the mean resultant length is obviously stationary (Fig 3A). By visual inspection, the average individual speed seems to decrease across time, and all time series are deemed nonstationary by a Kwiatkowski-Phillips-Schmidt-Shin test ($p < 0.01$), and so we may be on shaky ground to use methods that require stationarity. Additionally, since only 3 trials are available, the permutation test and the simultaneous permute-match test cannot detect dependence at the 0.05 level, so we perform a sequential permute-match test.

We arbitrarily started with average individual speed and identified a "speed"-perfect match, indicating a significant correlation between average individual speed and mean resultant length (Fig 3B; $p = 1/3^3 \approx 0.04$). Significance does not depend on the choice of which match is tested, since both a "speed"-perfect and a "mean resultant length"-perfect match are obtained. Note that although trials are independent of each other, all between-trial correlations from the full dataset are positive (off-diagonal elements in Fig 3B), illustrating the danger in applying naive data analysis methods to data that are autocorrelated or even nonstationary.

To try a parametric alternative, for each trial we time-averaged the average individual speed and mean resultant length over the full 10 minutes (Fig 3C; where each point is one trial). The sample Pearson correlation of the time-averaged variables is 0.99996, and a one-tailed test of significance (using the function stats.pearsonr from the Python package Scipy [36]) also detects dependence ($p \approx 0.003$), consistent with the permute-match tests. This test has the null hypothesis that the two variables shown in Fig 3C (time-averaged average individual speed and time-averaged mean resultant length) are independent and Gaussian. We view this test as comparable to the permute-match tests because although it requires a distributional assumption, it can be valid even when the time series are nonstationary.

Since data are limited in many scientific applications, we explored whether the results of the sequential permute-match tests and the parametric Pearson correlation test were consistent across different time series lengths. We truncated time series to various lengths (from 20 seconds to 10 minutes), and checked whether each test reported a significant ($p \leq 0.05$) correlation (Fig

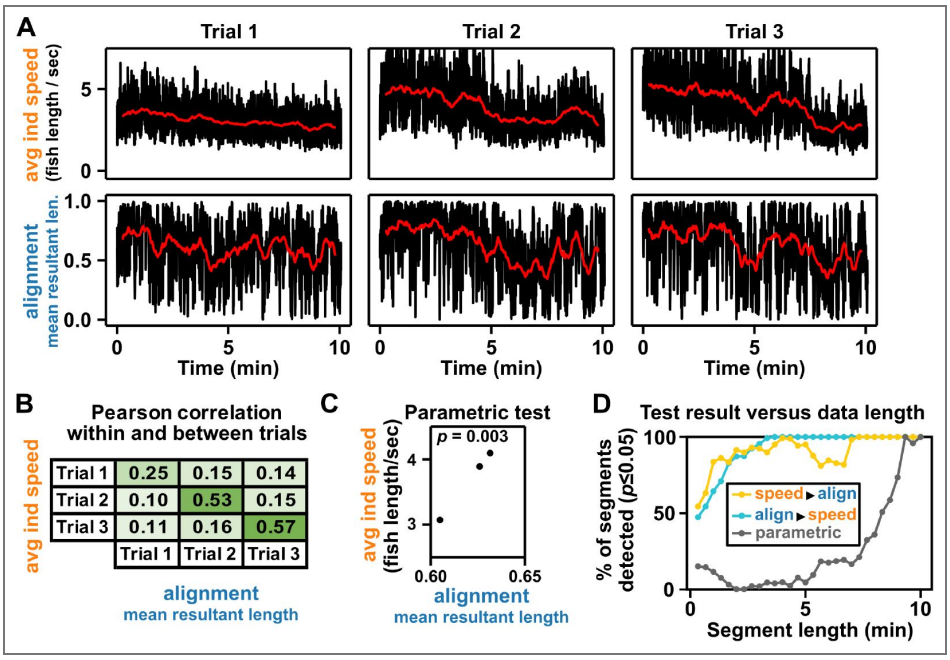


**Figure 3. A sequential permute-match test detects dependence between average individual speed and alignment (as measured by mean resultant length) in small groups of zebrafish.**

(**A**) Time series of average individual speed and mean resultant length for the three replicate videos. Black curves show original time series, and red curves show a 30-second moving average. Average individual speed appears to decrease with time in all trials. All 12 time series (2 variables × 3 trials × 2 smoothing conditions) were deemed non-stationary by a Kwiatkowski-Phillips-Schmidt-Shin (KPSS; [37]) test ($p < 0.01$). Note that the KPSS test seeks to reject a stationary null hypothesis in contrast to other common tests, whose null hypotheses are nonstationary. (**B**) Pearson correlation between alignment and average individual speed, both within trials and between trials, using all frames up to the length of the shortest speed and alignment time series (19308 frames; 10.06 minutes). Table entries are shaded by correlation. We observe a speed-perfect match (and an alignment-perfect match), and thus detect dependence with $p = 1/3^3 \approx 0.04$. (**C**) A parametric test also detects dependence. As a parametric alternative, for each trial we averaged values of speed and mean resultant length over the complete time series (visualized in the scatter plot shown, where each point is one trial). The sample Pearson correlation of the time-averaged variables is 0.99996, and a one-tailed test of significance gives $p \approx 0.003$. (**D**) Permute-match tests detected a significant correlation between speed and alignment more consistently than the parametric test. For a grid of lengths between 20 and 600 seconds we sampled 500 random segments of each length, each drawn from the first 600 seconds, and determined for each segment whether the parametric test and/or the two possible permute-match tests detected a significant ($p \leq 0.05$) correlation. In the edge case of the maximum 600-second length, all 500 "random" segments were identical. We used the Pearson correlation as the correlation function $\rho$ in the permute-match tests, and a one-tailed parametric test, reflecting the alternative hypothesis of correlation being positive (not merely nonzero).

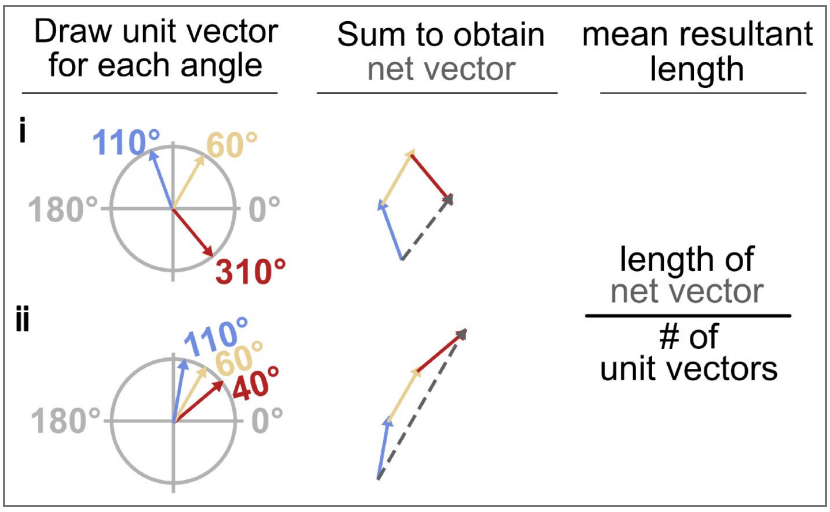


**Figure 4. The mean resultant length is a measure of alignment among angles.**

Consider three unit vectors colored blue, yellow, and red. The length of net vector (dashed) divided by the number of vectors is defined as the mean resultant length. Compared to less-aligned angles (i), strongly aligned angles (ii) produce a longer net vector, and thus a greater mean resultant length.

3D ). Both sequential permute-match tests reported a significant correlation at all lengths tested, whereas the parametric test was inconsistent, finding a significant correlation for about one third of lengths.

## Discussion

Outside the context of time series, permutation tests have been celebrated because they require only minimal assumptions [38]. To satisfy the requirements of permutation tests, it is sufficient (see Corollary 4 in Appendix 1) to collect data from exchangeable replicates. Data in each trial can be autocorrelated (e.g., temporally or spatially) and nonstationary. No distributional assumptions are required, and the correlation function $\rho$ can be completely arbitrary. For instance, $\rho$ could include a lag to detect delayed dependence, or even evaluate the correlation strength at several lags and report the strongest among them [39]. Moreover, the $\rho$ can be asymmetric so that $\rho(X, Y) \neq \rho(Y, X)$ as occurs when correlations are based on techniques that use one time series to predict values of another (e.g., Appendix 4). Such freedom from assumptions stands in contrast to the situation in the analysis of single-replicate time series, where the practitioner may be forced to make assumptions that are difficult to verify. For example, statistical tests of correlation between two time series often require stationary data at a minimum [40]. Yet, it is difficult to verify that a single time series is stationary because stationarity is a property of an entire ensemble of time series [40], so checking for stationarity in a single time series is similar in spirit to checking whether a single data point is drawn from a Gaussian distribution.

However, applying a trial-swapping permutation test to check for significant correlations requires that a sufficient number of trials be available to permute amongst one another. When only small numbers of trials are available, the permutation test is severely limited by mathematics. The minimum $p$-value that can possibly be achieved by a permutation test is $1/n!$, which will provide only modest evidence against a null hypothesis when $n = 4$ and is essentially useless when $n = 3$, regardless of how strong the true dependence may be. This limitation is particularly severe for researchers testing several related hypotheses, as this ideally requires a test that can produce $p$-values sufficiently low to maintain significance even after a multiple-testing correction.

Here we have described modified permutation tests – the sequential and simultaneous permute-match tests – that have access to lower $p$-values. In the sequential permute-match test, if the first variable does not display a perfect match, this test falls back to the simultaneous permute-match test. Thus, the sequential permute-match test is more powerful than the simultaneous test. However, the sequential test requires the arbitrary choice of checking for an $X$- or $Y$-perfect match first, making it potentially less reproducible between studies than the simultaneous test, which does not involve an arbitrary choice. If neither an $X$-nor $Y$-perfect match is obtained, both tests default to a permutation test. This step is valid because the tests are nested in the sense that a perfect match can only occur when the permutation $p$-value takes on its lowest possible value (Lemma 9 in Appendix 2).

The test appears computationally tractable for relevant sample sizes: Our implementation of the permute-match$_{(X \blacktriangleright Y)}$ procedure completed a single test of dependence in the setting of Fig 2 with an average runtime of 3 seconds when $n = 10$ on a 2023 14-inch MacBook Pro with an M2 Pro processor and 16 GB RAM (see Source data 1). For $n > 10$, a standard permutation test already can report a $p$-value below $3 \times 10^{-8}$, so the perfect match test is likely unnecessary for typical applications.

It seems likely that the tests could be further modified to report an even lower $p$-value when an $X$- and $Y$-perfect match occur simultaneously, as in Fig 3B. However, we have not investigated further and this problem is left for future efforts.

Although here we focused on time series, the permute-match test can in principle be applied to other scenarios where iid replicates are available, but data within each replicate are dependent. One example is spatial data in brain imaging, where the permutation test has been performed using multiple scanned brains [22]. Beyond spatial processes, dependent data also appear in

nucleotide sequences and natural language text. Overall, advances in methods that take advantage of multiple replicates may facilitate statistical testing without requiring assumptions that are difficult to verify.

# Methods

## Data preprocessing

We obtained fish trajectory data [34] from the web address at *https://drive.google.com/drive/folders/1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD*. For the first trial we used the file at the path "10/1/trajectories_wo_gaps.npy" (where "10" indicates the number of fish and "1" indicates the first trial). For the second and third trials, the file path was the same except it began with "10/2/" and "10/3/" respectively. These files contain sequences of fish positions indexed by video frame, as well as constants such as the frame rate (32 frames per second) and approximate fish body length.

We estimated the velocity of each fish as

$$\vec{v}_{i,t} = \frac{\vec{p}_{i,t+1} - \vec{p}_{i,t-1}}{2} R$$

where $\vec{p}_{i,t}$ is a 2-dimensional column vector specifying the position of the *i*th fish in units of body length at video frame *t*, $\vec{v}_{i,t}$ is a 2-dimensional column vector specifying the fish velocity in units of body length per second at video frame *t*, and *R* is the frame rate (32 frames per second). The 3 trials contained similar but slightly different numbers of frames. We truncated the $\vec{v}_{i,t}$ time series to the shortest among them (19308 frames ≈ 10.06 minutes) so they could be compared before downstream analyses.

A group's average individual speed at each time was then given by

$$\frac{1}{N} \sum_{i=1}^{N} \left| \vec{v}_{i,t} \right|$$

where *N* is the number of fish. The mean resultant length $C_t$ of a group at each time was given by the formula

$$C_t = \frac{1}{N} \left| \sum_{i=1}^{N} \frac{\vec{v}_{i,t}}{\left| \vec{v}_{i,t} \right|} \right| \tag{2}$$

Velocity vectors $\vec{v}_{i,t}$ covering missing tracking data (fewer than 0.02% of such vectors in any trial) were excluded from calculations of average individual speed and mean resultant length. When calculating mean resultant length, fish with speed of exactly zero (whose direction of motion is undefined) were also excluded. Instances of zero speed were rare, occurring in at most 1 fish-frame in any trial. Average individual speed and mean resultant length were determined from the remaining fish (always at least 8 of 10), so no time point was dropped from any trial.

## Statistical analysis

The permute-match test is described in Fig 1 and its validity is proved in Appendix 2. For the KPSS test, we used the function "statsmodels.tsa.stattools.kpss" in the Python package Statsmodels (ver. 0.10.1) [41]. We set the "regression" argument to "c" (which tests the null hypothesis of stationarity rather than trend-stationarity), and we set the "lags" argument to "auto" (which means that the number of lags used for testing stationarity is automatically chosen by a data-dependent method).

## Artificial intelligence

ChatGPT 5.2 was used while constructing the proof of Lemma 15, particularly in the application of the "taking out what is known" property. Claude Opus 4.8 was used to review code, mathematical arguments, and article text for errors and unclear language. Claude Opus was also used to edit

some parts of the manuscript for brevity. Corrections suggested by Claude Opus were reviewed by the authors and incorporated where appropriate.

## Data availability

Source code 1 contains all the simulation and analysis code used to prepare this manuscript. The experimental data analyzed in this paper can be obtained at the following web address: *https://drive.google.com/drive/folders/1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD*

## Acknowledgements

We thank Yunyi Zhang (CUHK-Shenzen) as well as members of the Shou Lab and two anonymous reviewers for valuable critiques. The University College London Computer Science High Performance Computational Cluster (*https://hpc.cs.ucl.ac.uk/*) provided computational resources and support that contributed to the completion of this research.

## Additional files

***Figure 2 - Source Data 1.*** Numerical values shown in Figure 2D.

***Figure 3 - Source Data 1.*** Numerical values shown in Figure 3D.

***Figure S2 - Source Data 1.*** Numerical values shown in Figure S2D.

***Figure S3 - Source Data 1.*** Numerical values shown in Figure S3.

***Source Code 1.*** Python scripts used to produce the results shown in Figures 2, 3, S2, and S3 as well as Source data 1.

***Source Data 1.*** Permute-match test run time as a function of the number of replicates.

## Appendix

In Appendix 1, we review the basic permutation test of independence and a proof of its validity. This is because, although these results are known to the statistical community, they are often presented in an abstract form whose connection to independence testing may not be obvious to the nonspecialist, or parts of the argument are simply left as an exercise to the reader. In Appendix 2, we provide the theoretical justification for the permute-match test, which is the main topic of this work. Appendix 3 expands on a brief claim made in the main text about *p*-values of sequential statistical tests, and Appendix 4 contains supplementary figures.

### 1 Justification of the permutation test of independence

In this section we provide a proof that the permutation test of independence is valid. No major novelty is claimed in this subsection. For instance, the results of Lemma 3 can be mostly reconstructed by piecing together various theorems, examples, and homework problems from section 15.2.1 of Lehmann and Romano [42]. Nevertheless, we present the complete argument here as a courtesy to the reader. Related proofs or proof sketches of various kinds of permutation tests can be found in [19, 29].

The following lemma describes a general rank-based method to test whether the distribution of a random variable changes upon applying a set of transformations. This lemma and proof are based on Theorem 15.2.1 in Lehmann and Romano [42]. We begin with this result because it comes first in the chain of logical steps that justify the permutation test, but it is somewhat abstract, and so the reader who prefers to start out with concrete concepts may wish to skip to Def 2 and return to this lemma when its necessity becomes apparent later. As a notational clarification, note that for the transformation $h_i$, to denote "$h_i$ applied to $Z$" we write $h_iZ$ rather than the more traditional $h_i(Z)$.

***Lemma 1.***

Let $Z \in \mathrm{Z}$ be a random variable. Let $H = \{h_1, \ldots, h_m\}$ be a set of functions from Z to Z such that:

1. the probability distributions of $Z$ and $h_iZ$ are the same for all functions $h_i$ in $H$, and
2. the unordered sets $\{h_1Z, \ldots, h_mZ\}$ and $\{h_1h_iZ, \ldots, h_mh_iZ\}$ are equal for all $h_i$ in $H$.

Let the statistic $T$ be a function from $\mathcal{Z}$ to $\mathbb{R}$. Given some $z \in \mathcal{Z}$, let the ordered values of $\{T(h_1z), \ldots, T(h_mz)\}$ be

$$T^{(1)}(z) \leq T^{(2)}(z) \leq \cdots \leq T^{(m)}(z)$$

Choose some significance level $\alpha \in [0, 1]$ and let the rank $k$ be

$$k = m - \lfloor m\alpha \rfloor \tag{S1}$$

where $\lfloor m\alpha \rfloor$ is the largest integer less than or equal to $m\alpha$. Then,

$$P\left(T(Z) > T^{(k)}(Z)\right) \leq \alpha$$

where $P(\cdot)$ denotes the probability of an event.

***Remark:***

Before proceeding to the proof, let us walk through the lemma, using a concrete example to help with the most abstract aspects. Suppose that random variable $Z$ is the result of a fair coin flip. That is, $Z = 0$ with 50% chance and $Z = 1$ with 50% chance. Then $\mathcal{Z}$ (the set of valid $Z$ values) is the set 0, 1, and $z$ can take on the value of 0 or 1. Next the lemma introduces $H$, which is a set of functions that need to satisfy certain requirements with respect to $Z$. Let's choose $H = h_1, h_2$ where $h_1$ "turns the coin over" and $h_2$ leaves the coin as it is. That is, $h_1z = 1 - z$ and $h_2z = z$. To check that $H$ satisfies requirement (1), note that flipping the fair coin over does not change the probability distribution of its outcome, and neither does leaving it alone. To see that $H$ satisfies requirement (2), we need to check that

$$\{h_1Z, h_2Z\} = \{h_1h_1Z, h_2h_1Z\} = \{h_1h_2Z, h_2h_2Z\}$$

If $Z = 0$, then the above expression becomes

$$\{1, 0\} = \{0, 1\} = \{1, 0\}$$

and if $Z = 1$, then it is

$$\{0, 1\} = \{1, 0\} = \{0, 1\}$$

and since these are unordered sets, they are all equal, so we have verified that $H$ satisfies requirement (2). Next, we need a function $T$. For example, choose $T(z) = z + 10$. Thus, for $z = 0$ we have $\{T(h_1z), T(h_2z)\} = \{11, 10\}$. The lemma then introduces a ranking notation where $T^{(i)}(Z)$ is the $i$th smallest term in the sequence. This notation is useful as we will be dealing with the set $\{T(h_1Z), \ldots, T(h_mZ)\}$ for potentially large values of $m$. Returning to our example, if $z = 0$, then $T^{(1)}(0) = 10$ and $T^{(2)}(0) = 11$. Similarly for $z = 1$, $T^{(1)}(1) = 10$ and $T^{(2)}(1) = 11$. Also, in the statement of the lemma, $T^{(i)}$ is defined as a function of $z$, not $Z$. This is to emphasize that the ordering depends on the particular outcome of the random variable and is not fixed. Finally, the lemma makes a claim about the probability that $T(Z)$ exceeds $T^{(k)}(Z)$, which is the $k$th lowest statistic in the set. In the example, if $\alpha = 0.05$, then $k = 2 - \lfloor 0.1 \rfloor = 2$. $T^{(2)}(Z)$ is always going to be the second smallest, i.e. the largest, element of $T(h_1z)$, $T(h_2z)$, which is 11. Of course, the probability of $T(Z)$ being *even greater* than the largest number is 0, and since 0 is less than $\alpha = 0.05$ we have verified the lemma in this case. As another example, suppose $\alpha = 0.5$. Then $k = 2 - \lfloor 1 \rfloor = 1$. $T^{(k)}(Z)$ is then $T^{(1)}(Z)$, the smallest element of $\{T(h_1z), T(h_2z)\}$. The probability of $T(Z)$ being greater than the smallest number is $0.5 \leq \alpha = 0.5$, which verifies the lemma for $\alpha = 0.5$.

***Proof of Lemma 1.***

Let $I(A)$ be the indicator function for the event $A$ (meaning that $I(A) = 1$ if $A$ occurs and $I(A) = 0$ otherwise).

First suppose that $T^{(k)}(Z)$ is not tied with any other $T^{(i)}(Z)$. Then Eq. S1 implies that there are $\lfloor m\alpha \rfloor$ values of $T^{(i)}(Z)$ that are greater than $T^{(k)}(Z)$. That is,

$$\begin{aligned}\lfloor m\alpha \rfloor &= \sum_{i=1}^{m} I\left(T\left(h_i Z\right) > T^{(k)}(Z)\right) \\ &= \sum_{i=1}^{m} I\left(T\left(h_i Z\right) > T^{(k)}\left(h_i Z\right)\right).\end{aligned}$$

To arrive at the second equality, we used the fact that $T^{(k)}(Z) = T^{(k)}(h_i Z)$, which follows from the second requirement of the lemma (i.e. $\{h_1 Z, \ldots, h_m Z\} = \{h_1 h_i Z, \ldots, h_m h_i Z\}$).

Alternatively, if there is possibly a tie for $T^{(k)}(Z)$ (meaning that there may be some $j \neq k$ such that $T^{(j)}(Z) = T^{(k)}(Z)$), then we have a similar formula, but with an inequality instead of an equality:

$$\lfloor m\alpha \rfloor \geq \sum_{i=1}^{m} I\left(T\left(h_i Z\right) > T^{(k)}\left(h_i Z\right)\right)$$

This version with the inequality is of course general since it is true under both cases. Using this formula,

$$m\alpha \geq \lfloor m\alpha \rfloor = E[\lfloor m\alpha \rfloor] \geq \sum_{i=1}^{m} E\left[I\left(T\left(h_i Z\right) > T^{(k)}\left(h_i Z\right)\right)\right].$$

Since $Z$ and $h_i Z$ have the same distribution, we may remove the $h_i$ in the above formula, and we have

$$m\alpha \geq \sum_{i=1}^{m} E\left[I\left(T(Z) > T^{(k)}(Z)\right)\right] = mE\left[I\left(T(Z) > T^{(k)}(Z)\right)\right].$$

Dividing through by $m$ gives

$$\alpha \geq E\left[I\left(T(Z) > T^{(k)}(Z)\right)\right] = P\left(T(Z) > T^{(k)}(Z)\right)$$

as required.

Next we give a definition of the permutation test. The definition here is somewhat more general than in the main text. As with $h_i$, we will denote "$g_i$ applied to $Z$" by writing $g_i Z$.

***Definition 2.***

*The permutation test of independence*

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be sequences of random variables. Let $G = \{g_1, \ldots, g_{n!}\}$ be the complete set of functions that permute a sequence of length $n$. For example, if $\boldsymbol{Y}$ is (2, 7, 4), then $g_i \boldsymbol{Y}$ might be (2, 4, 7). Define the statistic $T(\boldsymbol{X}, \boldsymbol{Y})$ to be a function that returns a real number, typically interpreted as the overall correlation strength. Repeatedly compute the statistic after permuting the elements of $\boldsymbol{Y}$. That is, compute

$$\left\{T\left(\boldsymbol{X}, g_1 \boldsymbol{Y}\right), T\left(\boldsymbol{X}, g_2 \boldsymbol{Y}\right), \ldots, T\left(\boldsymbol{X}, g_{n!} \boldsymbol{Y}\right)\right\}$$

Let $N_{\geq}$ be the number of permuted statistic values that are at least as large as the original statistic value. That is,

$$N_{\geq} = \sum_{i=1}^{n!} I\left(T\left(\boldsymbol{X}, g_i \boldsymbol{Y}\right) \geq T(\boldsymbol{X}, \boldsymbol{Y})\right)$$

where $I(A)$ is the indicator function for the event $A$ (meaning that $I(A) = 1$ if $A$ occurs and $I(A) = 0$ otherwise). Then the $p$-value of the test is given by

$$p_{\text{perm}} = N_{\geq}/n!$$

We now prove the validity of the permutation test.

***Lemma 3.***

*The permutation test of independence is valid when* $\boldsymbol{Y}$ *is exchangeable.*

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be sequences of random variables such that $\boldsymbol{Y}$ is exchangeable and where $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$. Perform the permutation test of independence (Def. 2) to obtain $p_{perm}$. Then, $P(p_{perm} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$.

**Proof:** Here we use the symbols $g_i$ and $T$ with their meanings from Def. 2.

We first set up the problem in a way that allows us to apply Lemma 1. To construct the random variable $Z$ and transformation set $H$ for Lemma 1, we choose $Z = (\boldsymbol{X}, \boldsymbol{Y})$ and $h_i Z = (\boldsymbol{X}, g_i \boldsymbol{Y})$. Lemma 1 has two requirements. First, $Z$ and $h_i Z$ must have the same distribution. This requirement is satisfied: Since $\boldsymbol{Y}$ is exchangeable and $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$, it follows that $(\boldsymbol{X}, \boldsymbol{Y})$ has the same distribution as $(\boldsymbol{X}, g_i \boldsymbol{Y})$. The second requirement is that $\{h_1 Z, \ldots, h_m Z\} = \{h_1 h_i Z, \ldots, h_m h_i Z\}$. This requirement is also satisfied: The set of permutations of $\boldsymbol{Y}$ is the same as the set of permutations of $g_i \boldsymbol{Y}$. Since both requirements are satisfied, we may apply Lemma 1.

We now prove the result directly. Below, $A \leftrightarrow B$ means "$A$ if and only if $B$".

$$
\begin{aligned}
p_{\text{perm}} \leq \alpha &\leftrightarrow N_{\geq} \leq n!\alpha \leftrightarrow \sum_{i=1}^{n!} I\left(T\left(\boldsymbol{X}, g_i \boldsymbol{Y}\right) \geq T(\boldsymbol{X}, \boldsymbol{Y})\right) \leq n!\alpha \\
&\leftrightarrow \sum_{i=1}^{n!} \left(1 - I\left(T\left(\boldsymbol{X}, g_i \boldsymbol{Y}\right) \geq T(\boldsymbol{X}, \boldsymbol{Y})\right)\right) \geq n! - n!\alpha \\
&\leftrightarrow \sum_{i=1}^{n!} \left(I\left(T\left(\boldsymbol{X}, g_i \boldsymbol{Y}\right) < T(\boldsymbol{X}, \boldsymbol{Y})\right)\right) \geq n! - n!\alpha \\
&\leftrightarrow \sum_{i=1}^{n!} \left(I\left(T\left(\boldsymbol{X}, g_i \boldsymbol{Y}\right) < T(\boldsymbol{X}, \boldsymbol{Y})\right)\right) \geq n! - \lfloor n!\alpha \rfloor .
\end{aligned}
$$

As in Eq. S1, let us define $k = m - \lfloor m\alpha \rfloor$, where $n! = m$. Then, the final line above says "the number of permuted $\boldsymbol{Y}$ s that produce a smaller $T$ than the original $\boldsymbol{Y}$ is greater than or equal to $k$." Said another way, "when we rank the $T$ s from least to greatest, the $T$ from the original $\boldsymbol{Y}$ will have a higher rank than $k$". This in turn is equivalent to $T(\boldsymbol{X}, \boldsymbol{Y}) > T^{(k)}(\boldsymbol{X}, \boldsymbol{Y})$. Thus, we may directly apply Lemma 1, thereby obtaining

$$
P\left(p_{\text{perm}} \leq \alpha\right) = P\left(T(\boldsymbol{X}, \boldsymbol{Y}) > T^{(k)}(\boldsymbol{X}, \boldsymbol{Y})\right) \leq \alpha
$$

as required.

Lemma 3 is "asymmetric" in that it requires $\boldsymbol{Y}$ to be exchangeable and says nothing about $\boldsymbol{X}$. It is possible to write down a variant of this lemma where either an exchangeable $\boldsymbol{Y}$ or an exchangeable $\boldsymbol{X}$ is sufficient, but this requires an extra condition on $T$. Specifically, it is required that applying the same permutation to both $\boldsymbol{X}$ and $\boldsymbol{Y}$ must leave the value of $T(\boldsymbol{X}, \boldsymbol{Y})$ unchanged. That is, $T(\boldsymbol{X}, \boldsymbol{Y}) = T(g_i \boldsymbol{X}, g_i \boldsymbol{Y})$ for any permutation function $g_i$. Importantly, the form of $T$ that is used for the permute-match test (i.e. $T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho\left(X_i, Y_i\right)$) satisfies this requirement because in this case, $T(g_i \boldsymbol{X}, g_i \boldsymbol{Y})$ simply rearranges the order of the terms in the summation, which clearly has no effect on the value of $T$. We establish this formally in the corollary below.

***Corollary 4.***

*If T is "nice", then the permutation test is valid when either* $\boldsymbol{X}$ *or* $\boldsymbol{Y}$ *is exchangeable.*

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be sequences of random variables such that at least one of $(\boldsymbol{X}, \boldsymbol{Y})$ is exchangeable and where $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$. Perform the permutation test of independence (Def. 2) using a correlation function with the "nice" property that

$$
T(\boldsymbol{X}, \boldsymbol{Y}) = T\left(g_i \boldsymbol{X}, g_i \boldsymbol{Y}\right) \tag{S2}
$$

for any permutation function $g_i$. Then, $P(p_{perm} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$.

**Proof:** Lemma 3 already covers the case where $\boldsymbol{Y}$ is exchangeable. We now consider the case where $\boldsymbol{X}$ is exchangeable, but not $\boldsymbol{Y}$.

We proceed by showing that the permutation test where $\boldsymbol{Y}$ is permuted is equivalent to the permutation test where $\boldsymbol{X}$ is permuted. Notice that to show this equivalence it is sufficient to establish that

$$\left\{T\left(\boldsymbol{X}, g_1\boldsymbol{Y}\right), \ldots, T\left(\boldsymbol{X}, g_{n!}^{Y}\right)\right\} = \left\{T\left(g_1\boldsymbol{X}, \boldsymbol{Y}\right), \ldots, T\left(g_{n!}\boldsymbol{X}, \boldsymbol{Y}\right)\right\}$$

where the left and right sides both denote unordered sets. It will be useful to refer to the inverse permutation $g_i^{-1}$, which is the permutation that "undoes" $g_i$. More precisely, $g_i^{-1}$ is defined by the relation $g_i^{-1}g_iZ = Z$. Using Eq. S2, we have:

$$\begin{aligned} T\left(\boldsymbol{X}, g_i\boldsymbol{Y}\right) &= T\left(g_i^{-1}\boldsymbol{X}, g_i^{-1}g_i\boldsymbol{Y}\right) \\ &= T\left(g_i^{-1}\boldsymbol{X}, \boldsymbol{Y}\right). \end{aligned} \tag{S3}$$

Note that the inverse permutation $g_i^{-1}$ is guaranteed to exist because any permutation $g_i$ can be "undone" by some other permutation $\left(g_i^{-1}\right)$. In the edge case where $g_i$ is the trivial identity permutation, $g_i^{-1}$ is also the identity permutation. Additionally, the set of permutation functions is the same as the set of inverse permutation functions, meaning that

$$\left\{g_1, \ldots, g_{n!}\right\} = \left\{g_1^{-1}, \ldots, g_{n!}^{-1}\right\} \tag{S4}$$

because each permutation is an inverse permutation (because $g_i$ inverts $g_i^{-1}$), and each inverse permutation is a permutation. It follows that

$$\begin{aligned} \left\{T\left(\boldsymbol{X}, g_1\boldsymbol{Y}\right), \ldots, T\left(\boldsymbol{X}, g_{n!}\boldsymbol{Y}\right)\right\} &= \left\{T\left(g_1^{-1}\boldsymbol{X}, \boldsymbol{Y}\right), \ldots, T\left(g_{n!}^{-1}\boldsymbol{X}, \boldsymbol{Y}\right)\right\} \\ &= \left\{T\left(g_1\boldsymbol{X}, \boldsymbol{Y}\right), \ldots, T\left(g_{n!}\boldsymbol{X}, \boldsymbol{Y}\right)\right\} \end{aligned}$$

where the first equality follows from Eq. S3 and the second equality follows from Eq. S4. This shows that the permutation test where $\boldsymbol{Y}$ is permuted is equivalent to the permutation test where $\boldsymbol{X}$ is permuted.

From Def. 2 and Lemma 3 it is clear that a permutation test where $\boldsymbol{X}$ is permuted instead of $\boldsymbol{Y}$ would be valid (i.e. $P\left(p_{perm} \leq \alpha\right) \leq \alpha$) if $\boldsymbol{X}$ is exchangeable. Since we have just now shown that the present procedure is equivalent to a permutation test where $\boldsymbol{X}$ is permuted instead of $\boldsymbol{Y}$, it follows that this procedure is valid when $\boldsymbol{X}$ is exchangeable, which completes the proof.

## 2 Validity of the permute-match test

We now justify the validity of the permute-match test, which is the primary contribution of the present work. Recall from the main text that the null hypothesis of this test is that $X_1$, ..., $X_n$ are iid, $Y_1$, ..., $Y_n$ are iid and $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$.

***Definition 5.***

*The perfect match events ($M_X$ and $M_Y$)*

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be sequences of random variables. Let $\rho$ (the "correlation function") be a function that maps a pair $(X_i, Y_j)$ to a real number. We say that the event $M_Y$ (also called a "*Y*-perfect match") occurs if and only if:

$$\rho(X_i, Y_i) > \rho(X_j, Y_i)$$

for all pairs $(i, j)$ such that $i \neq j$.

Similarly, the event $M_X$ (also called an "*X*-perfect match") occurs if and only if

$$\rho(X_i, Y_i) > \rho(X_i, Y_j)$$

for all pairs $(i, j)$ such that $i \neq j$.

***Lemma 6.***

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sequence of random variables and let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a sequence of iid random variables. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Then, the probability of a *Y*-perfect match is at most $1/n^n$.

**Proof:** Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be the sample spaces of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (i.e. $\boldsymbol{X} \in \boldsymbol{X}$ and $\boldsymbol{Y} \in \boldsymbol{Y}$). Let $\boldsymbol{c}$ be a function that maps from $\boldsymbol{X} \times \boldsymbol{Y}$ to $\mathbb{N}^n$. Specifically, $\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) = (c_1(\boldsymbol{x}, \boldsymbol{y}), \ldots, c_n(\boldsymbol{x}, \boldsymbol{y}))$, where each $c_i$ is given as follows:

$$c_i(\boldsymbol{x}, \boldsymbol{y}) = \underset{j}{\operatorname{argmax}} \left(\rho(x_j, y_i)\right)$$

That is, $c_i(\boldsymbol{x}, \boldsymbol{y}) \in \{1, \ldots, n\}$ is defined to be the index $j$ that maximizes $\rho(x_j, y_i)$. In the case of a multi-way tie, $c_i$ is the lowest from among the maximizing index values. (For example, if $j = 3$ and $j = 7$ both produced the greatest value of $\rho(x_j, y_i)$, then $c_i$ would be 3 since $3 < 7$.) Note that if a *Y*-perfect match occurs, then

$$c_1(\boldsymbol{X}, \boldsymbol{Y}) = 1, \ldots, c_n(\boldsymbol{X}, \boldsymbol{Y}) = n$$

Consider $\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{Y})$ for a fixed $\boldsymbol{x} \in \boldsymbol{X}$. Notice that

$$c_1(\boldsymbol{x}, \boldsymbol{Y}), \ldots, c_n(\boldsymbol{x}, \boldsymbol{Y})$$

are iid because each $c_i(\boldsymbol{x}, \boldsymbol{Y})$ depends only on $Y_i$. Since the $c_i(\boldsymbol{x}, \boldsymbol{Y})$ terms are identically distributed across $i$, we have:

$$P(c_i(\boldsymbol{x}, \boldsymbol{Y}) = i) = P(c_1(\boldsymbol{x}, \boldsymbol{Y}) = i)$$

for all $i$. Therefore we have

$$\sum_{i=1}^{n} P(c_i(\boldsymbol{x}, \boldsymbol{Y}) = i) = \sum_{i=1}^{n} P(c_1(\boldsymbol{x}, \boldsymbol{Y}) = i) = 1$$

By dividing through by $n$, we may write

$$\frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} P(c_i(\boldsymbol{x}, \boldsymbol{Y}) = i) \geq \left( \prod_{i=1}^{n} P(c_i(\boldsymbol{x}, \boldsymbol{Y}) = i) \right)^{1/n} \tag{S5}$$

where the inequality comes from the AM-GM inequality (e.g., [43] pg. 20). Since the $c_i(\boldsymbol{x}, \boldsymbol{Y})$ terms are independent,

$$\prod_{i=1}^{n} P(c_i(\boldsymbol{x}, \boldsymbol{Y}) = i) = P(c_1(\boldsymbol{x}, \boldsymbol{Y}) = 1, \ldots, c_n(\boldsymbol{x}, \boldsymbol{Y}) = n)$$

Substituting this fact into inequality S5 and raising both sides to the $n$th power gives us

$$\frac{1}{n^n} \geq P(c_1(\boldsymbol{x}, \boldsymbol{Y}) = 1, \ldots, c_n(\boldsymbol{x}, \boldsymbol{Y}) = n) = P(\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{Y}) = [1, \ldots, n]) \tag{S6}$$

Since the above inequality holds for all $\boldsymbol{x} \in \boldsymbol{X}$, and since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, Lemma 7 (given immediately following this proof) implies

$$P(\boldsymbol{c}(\boldsymbol{X}, \boldsymbol{Y}) = [1, \ldots, n]) \leq \frac{1}{n^n} \tag{S7}$$

Since a $Y$-perfect match implies $\boldsymbol{c}(\boldsymbol{X}, \boldsymbol{Y}) = [1, \ldots, n]$, the left side of this inequality has a lower bound of $P(M_Y)$, giving us

$$P(M_Y) \leq P(\boldsymbol{c}(\boldsymbol{X}, \boldsymbol{Y}) = [1, \ldots, n]) \leq \frac{1}{n^n}$$

as required.

We now justify the step from Equation S6 to Equation S7 in the proof above. That is, we explain why a bound that holds for every fixed $x$ continues to hold when $x$ is replaced by the random variable $X$, provided $X$ and $Y$ are independent.

***Lemma 7.***

Let $X \in X$ and $Y \in Y$ be independent random variables. Let $f$ be a function whose domain is the product space $X \times Y$ and let $v$ be a member of the range of $f$. Assume that there exists a constant $p \in [0, 1]$ such that

$$P(f(x, Y) = v) \leq p \text{ for all } x \in \mathcal{X}$$

Then,

$$P(f(X, Y) = v) \leq p$$

**Proof:** Let $I_v(x, y)$ denote the indicator function for the event where $f(x, y) = v$. That is, $I_v(x, y) = 1$ if $f(x, y) = v$ and otherwise $I_v(x, y) = 0$.

From the law of total expectation we have

$$\begin{aligned} P(f(X, Y) = v) &= E[I_v(X, Y)] \\ &= E[E(I_v(X, Y) \mid X)] \end{aligned} \tag{S8}$$

Let $g(x)$ be a function given by

$$g(x) = E[I_v(x, Y)]$$

Since $X$ and $Y$ are independent, it follows (e.g., see Example 4.1.7 in [44]) that $g(X) = E(I_v(X, Y) X)$. Substituting this fact into equation S8 gives

$$P(f(X, Y) = v) = E[g(X)].$$

Note that it follows from the conditions of this lemma that $g(x) \leq p$ for all $x \in$ X. Thus, monotonicity of the expected value implies $E[g(X)] \leq E[p] = p$. Combining this fact with the above equation, we obtain

$$P(f(X, Y) = v) \leq p$$

as required.

***Lemma 8.***

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sequence of iid random variables and let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a sequence of random variables. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Then, the probability of an $X$-perfect match is at most $1/n^n$.

**Proof:** The argument is analogous to that of Lemma 6. The main differences are that in this case we define $c_i(\boldsymbol{x}, \boldsymbol{y}) = \underset{j}{\operatorname{argmax}} (\rho(x_i, y_j))$ instead of $c_i(\boldsymbol{x}, \boldsymbol{y}) = \operatorname{argmax}(\rho(x_j, y_i))$, and here we condition on $\boldsymbol{Y} = \boldsymbol{y}$ instead of $\boldsymbol{X} = \boldsymbol{x}$.

***Lemma 9.***

*A perfect match implies* $p_{perm} = 1/n!$

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be sequences of random variables. Let $\rho$ be a function that maps a pair $(X_i, Y_j)$ to a real number. Define the average correlation

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i). \tag{S9}$$

Perform the permutation test of independence (Def. 2) using $T$ as the statistic, and obtain the $p$-value $p_{perm}$. Also check for an $X$-perfect and $Y$-perfect match test (Def. 5) using $\rho$ as the correlation function. Then, the occurrence of either an $X$-perfect match or a $Y$-perfect match implies that $p_{perm} = 1/n!$.

**Proof:** This claim is shown by a visual argument in Fig 1C. Here we walk through the same logic in prose.

Suppose a $Y$-perfect match occurs, meaning that

$$\rho(X_i, Y_i) > \rho(X_j, Y_i) \; \forall (i \neq j)$$

In other words, for each $Y_i$, its correlation term $\rho(\cdot, Y_i)$ is strictly maximized when $Y_i$ is paired with $X_i$. Applying a permutation to $\boldsymbol{Y}$ (other than the trivial identity permutation) means that for at least one choice of $i$, the $\rho(X_i, Y_i)$ term in Eq. S9 becomes replaced with $\rho(X_j, Y_i)$ for some $j \neq i$. But $\rho(X_i, Y_i) > \rho(X_j, Y_i)$. Therefore, $T(\boldsymbol{X}, \boldsymbol{Y})$ must be strictly greater than all permuted versions $T(\boldsymbol{X}, g_i\boldsymbol{Y})$ (other than when $g_i$ is the identity permutation). It follows that $p_{perm} = 1/n!$.

Suppose now that an $X$-perfect match occurs. It follows that for each $X_i$ its correlation term $\rho(X_i, \cdot)$ is strictly maximized when $X_i$ is paired with $Y_i$. Applying a permutation to $\boldsymbol{X}$ (other than the identity permutation) means that for at least one choice of $i$, the $\rho(X_i, Y_i)$ term in Eq. S9 becomes replaced with $\rho(X_i, Y_j)$ for some $j \neq i$. Therefore, $T(\boldsymbol{X}, \boldsymbol{Y})$ is strictly greater than all permuted versions $T(g_i\boldsymbol{X}, \boldsymbol{Y})$ whenever $g_i$ is not the identity permutation. So $p_{perm} = 1/n!$.

***Lemma 10.***

$1/n! \geq 2/n^n$ for all integers $n > 1$, with equality only when $n = 2$.

**Proof:** When $n = 2$ it is clear that $1/n! = 2/n^n = 1/2$.

For $n > 2$, we establish $1/n! > 2/n^n$ by an inductive argument. For the base case, suppose $n = 3$. Then, $1/n! > 2/n^n$ becomes $1/6 > 2/27$, which is true.

For the inductive step, suppose the target inequality $1/n! > 2/n^n$ holds. Multiplying the target inequality on both sides by $1/(n+1)$, we obtain

$$\frac{1}{(n+1)!} > \frac{2}{n^n}\left(\frac{1}{n+1}\right) > \frac{2}{(n+1)^n}\left(\frac{1}{n+1}\right) = \frac{2}{(n+1)^{n+1}}$$

This shows that $1/n! > 2/n^n$ implies $1/(n+1)! > 2/(n+1)^{n+1}$. We conclude by the inductive principle that the target inequality is true for $n > 2$.

***Theorem 11.***

*The permute-match*$_{X \blacktriangleright Y}$ *test is valid.*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences of iid random variables. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Let the correlation function $\rho$ be a function that maps a pair $(X_i, Y_j)$ to a real number. Define the average correlation

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho\left(X_i, Y_i\right)$$

Perform the permutation test of independence (Def. 2) using $T$ as the statistic, and obtain the $p$-value $p_{perm}$. Also check for an $X$-perfect match test and a $Y$-perfect match test (Def. 5) using $\rho$ as the correlation function. Let

$$p = \begin{cases} 1/n^n & \text{if an } X\text{-perfect match occurs} \\ 2/n^n & \text{if a } Y\text{-perfect match occurs, but not an } X\text{-perfect match} \\ p_{\text{perm}} & \text{if neither occurs} \end{cases}$$

Then $P\,(p \le \alpha) \le \alpha$ for $\alpha \in [0, 1]$ and $n > 1$.

**Proof:** For notational convenience, we denote the events of an $X$-perfect match and $Y$-perfect match as $M_X$ and $M_Y$.

There are four cases to consider. Either (case i) $\alpha \ge 1/n!$, or (case ii) $1/n! > \alpha \ge 2/n^n$, or (case iii) $2/n^n > \alpha \ge 1/n^n$, or (case iv) $\alpha < 1/n^n$. Note that Lemma 10 tells us that $1/n! \ge 2/n^n$.

Case i: Suppose $\alpha \ge 1/n!$. In this case, $p \le \alpha$ if $M_X$ occurs or $M_Y$ occurs or $p_{perm} \le \alpha$.

$$\begin{aligned} P(p \le \alpha) &= P\left(p_{\text{perm}} \le \alpha \text{ or } M_X \text{ or } M_Y\right) \\ &= P\left(p_{\text{perm}} \le \alpha \text{ or } \left(M_X \text{ or } M_Y\right)\right) \end{aligned}$$

This may be split into 3 terms via the identity $P\,(A \text{ or } B) = P\,(A \text{ and } B) + P\,(A \text{ and not } B) + P\,((\text{not } A) \text{ and } B)$, as follows:

$$\begin{aligned} P(p \le \alpha) = {} & P\left(p_{\text{perm}} \le \alpha \text{ and } \left(M_X \text{ or } M_Y\right)\right) \\ & + P\left(p_{\text{perm}} \le \alpha \text{ and not } \left(M_X \text{ or } M_Y\right)\right) \\ & + P\left(p_{\text{perm}} > \alpha \text{ and } \left(M_X \text{ or } M_Y\right)\right) \end{aligned}$$

However, by Lemma 9, if either $M_X$ or $M_Y$ occurs, then $p_{perm} = 1/n!$, which would in turn imply $p_{perm} \le \alpha$ under case i. That is, the term $P\,(p_{perm} > \alpha \text{ and } (M_X \text{ or } M_Y))$ in the above summation is zero. Removing this term, we have

$$\begin{aligned} P(p \le \alpha) &= P\left(p_{\text{perm}} \le \alpha \text{ and } \left(M_X \text{ or } M_Y\right)\right) + P\left(p_{\text{perm}} \le \alpha \text{ and not } \left(M_X \text{ or } M_Y\right)\right) \\ &= P\left(p_{\text{perm}} \le \alpha\right) \le \alpha \end{aligned}$$

where the final inequality is due to Lemma 3. This establishes the claim for case i.

Case ii: Suppose $1/n! > \alpha \ge 2/n^n$. In this case, $p_{perm} > \alpha$ since $p_{perm}$ cannot be less than $1/n!$, so we have

$$\begin{aligned} P(p \le \alpha) &= P\left(M_X \text{ or } M_Y\right) \\ &\le P\left(M_X\right) + P\left(M_Y\right) \end{aligned}$$

From Lemmas 6-8 we know $P\,(M_X) \le 1/n^n$ and $P\,(M_Y) \le 1/n^n$. Thus the above inequality becomes

$$P(p \le \alpha) \le P\left(M_X\right) + P\left(M_Y\right) \le 1/n^n + 1/n^n \le \alpha$$

which establishes the claim in case ii.

Case iii: Suppose $2/n^n > \alpha \ge 1/n^n$. In this case, only $M_X$ will give us $p \le \alpha$. So we have

$$P(p \le \alpha) = P\left(M_X\right) \le 1/n^n \le \alpha$$

which verifies the claim in case iii.

Case iv: If $\alpha < 1/n^n$, then, $p \le \alpha$ is impossible so $P\,(p \le \alpha) = 0 \le \alpha$, so the claim is trivially satisfied in case iv.

Having been shown in all four cases, the claim has been proven.

***Theorem 12.***

*The permute-match$_{Y \blacktriangleright X}$ test is valid.*

Consider the same setting as Theorem 11, except that now

$$p = \begin{cases} 1/n^n & \text{if a } Y\text{-perfect match occurs} \\ 2/n^n & \text{if an } X\text{-perfect match occurs, but not a } Y\text{-perfect match} \\ p_{\text{perm}} & \text{if neither occurs} \end{cases}$$

Then, $P(p \le \alpha) \le \alpha$ for $\alpha \in [0, 1]$ and $n > 1$.

**Proof:** The proof is analogous to that of Theorem 11. In fact, all we need to change here is that in case iii, we use the fact that $P(M_Y) \le 1/n^n$ instead of that $P(M_X) \le 1/n^n$.

***Theorem 13.***

*The simultaneous permute-match$_{(X;Y)}$ test is valid.*

Consider the same setting as Theorem 11, except that now

$$q = \begin{cases} 2/n^n & \text{if an } X\text{-perfect match or } Y\text{-perfect match occurs} \\ p_{\text{perm}} & \text{if neither occurs} \end{cases}$$

Then, $P(q \le \alpha) \le \alpha$ for $\alpha \in [0, 1]$ and $n > 1$.

**Proof:** We first show by cases that $p$ of Theorem 11 is always less than or equal to $q$. First, if an $X$-perfect match occurs, then $p = 1/n^n < 2/n^n = q$. Second, if there is a $Y$-perfect match but not an $X$-perfect match, then $p = 2/n^n = q$. Third, if neither an $X$-perfect nor a $Y$-perfect match occurs, then $p = p_{perm} = q$. Overall, $p \le q$. It follows that $q \le \alpha$ implies $p \le \alpha$.

Then it follows from Theorem 11 that

$$P(q \le \alpha) \le P(p \le \alpha) \le \alpha$$

so $P(q \le \alpha) \le \alpha$ as required.

We have seen that $1/n^n$ is a universal upper bound on $P(X - \text{perfect match})$. Here, we give an intuitive explanation for why this probability can be far smaller than $1/n^n$, and also why it is still possible to construct examples where $P(X - \text{perfect match})$ comes arbitrarily close to $1/n^n$.

To see why $P(X - \text{perfect match})$ can fall well below $1/n^n$, recall the analogy of students and advisors, where $\rho(X_i, Y_j)$ measures how well a student $i$ likes advisor $j$ and an $X$-perfect match occurs if each student prefers their own advisor over all the others. Suppose that all the students value essentially the same qualities in an advisor. In the most extreme instance, $\rho(X_i, Y_j)$ depends only on $Y_j$ and not $X_i$ at all. Then all students will rank advisors in the same way, so they will all prefer the same advisor, making an $X$-perfect match impossible.

This suggests that to maximize the chance of an $X$-perfect match, different students must care about different "dimensions" of their advisors. However, under the null hypothesis, the $X_i$ variables are independent and identically distributed, so we cannot deterministically assign different advisor dimensions to different students. Instead, the key trick in our construction is to introduce an enormous number of possible dimensions. By allowing the number of advisor dimensions ($k$ in the following formal construction) to be arbitrarily large, we make the probability that two students happen to care about the same dimension arbitrarily small. And, if indeed each student cares about a different (and independent) dimension, all the students will rank advisors independently, so that the probability of an $X$-perfect match approaches $1/n^n$.

The following formalizes the above intuition, establishing a scenario where under the null hypothesis, the probability of an $X$-perfect match can be made arbitrarily close to $1/n^n$. An analogous example could be constructed for a $Y$-perfect match.

***Construction.***

Let $k$ be a positive integer whose value will be chosen later. For each $i \in \{1, ..., n\}$, let $X_i$ be drawn uniformly at random from among $\{1, ..., k\}$, and let $X_1, ..., X_n$ be independent.

For each $j \in \{1, ..., n\}$ and each $l \in \{1, ..., k\}$, let

$$Y_{j,l} \sim \text{Uniform}(0, 1)$$

where all $Y_{j,l}$ terms are independent of each other and of the $X_i$ terms. Define $Y_j = [Y_{j,1}, \ldots, Y_{j,k}]$. In the story above, $X_i$ is the dimension that student $i$ cares about, and $Y_{j,l}$ is the quality of advisor $j$ along dimension $l$.

Let the correlation function be

$$\rho(X_i, Y_j) = Y_{j,X_i}$$

In this case, we say that an $X$-perfect match occurs if and only if

$$Y_{i,X_i} > Y_{j,X_i} \text{ for all pairs } (i,j) \text{ such that } j \neq i$$

***Proposition 14.***

Let $q$ be a number such that $0 \leq q < 1$. Then, for the construction above, when $H_0$ is satisfied (i.e., the $X_i$ are iid, the $Y_i$ are iid, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent), the probability of an $X$-perfect match is $>q/n^n$.

**Proof:** Fix $x \in \{1, \ldots, k\}$ and let $A_{j,x}$ be the event that $Y_{j,x}$ is the strictly largest element from the sequence $Y_{[:,x]} = [Y_{1,x}, \ldots, Y_{n,x}]$. Since the elements of $Y_{[:,x]}$ are iid continuous random variables, the probability of a "tie" for largest is zero, and so it follows from symmetry that $P(A_{j,x}) = 1/n$ for all $j$. Now, fix $\{x_1, \ldots, x_n\}$ where each $x_i$ is unique. Then, $A_{i,x_i}$ are all independent of one another because each $A_{i,x_i}$ depends only on $Y_{[:,x_i]}$, and the $Y_{[:,x_i]}$ are independent. Thus

$$P\left(\bigcap_{i=1}^n A_{i,x_i}\right) = \prod_{i=1}^n P(A_{i,x_i}) = 1/n^n$$

Notice that for $\boldsymbol{X} = \boldsymbol{x}$, an $X$-perfect match occurs if and only if all the $A_{i,x_i}$ events occur. Let us formalize this further. Let $\mathrm{M}_X$ be a set in the product sample space of $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that $(\boldsymbol{x}, \boldsymbol{y}) \in \mathrm{M}_X$ if and only if $y_{i,x_i} > y_{j,x_i}$ for all pairs $i \neq j$. Moreover, let us denote an $X$-perfect match by the symbol $M_X$. Notice that $M_X$ occurs if and only if $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathrm{M}_X$. Now, we may rewrite the above equation as

$$P((\boldsymbol{x}, \boldsymbol{Y}) \in \mathscr{M}_X) = 1/n^n$$

for all $\boldsymbol{x}$ whose entries are unique. Applying Lemma 15 (given just after the present proof) to the above result, we see that

$$P((\boldsymbol{X}, \boldsymbol{Y}) \in \mathscr{M}_X \mid X_1 \neq \cdots \neq X_n) = 1/n^n$$

We are now ready to complete the result. We start with

$$P((\boldsymbol{X}, \boldsymbol{Y}) \in \mathscr{M}_X) \geq P((X_1 \neq \cdots \neq X_n) \bigcap (\boldsymbol{X}, \boldsymbol{Y}) \in \mathscr{M}_X)$$

Substituting $M_X \Longleftrightarrow (\boldsymbol{X}, \boldsymbol{Y}) \in \mathrm{M}_X$, we now obtain:

$$P(M_X) \geq P\left((X_1 \neq \cdots \neq X_n) \bigcap M_X\right) \tag{S10}$$

$$P(M_X) \geq P(X_1 \neq \cdots \neq X_n)\, P(M_X \mid X_1 \neq \cdots \neq X_n) \tag{S11}$$

$$P(M_X) \geq P(X_1 \neq \cdots \neq X_n)/n^n \tag{S12}$$

Since $X_1, \ldots, X_n$ are uniformly distributed, we have

$$P(X_1 \neq \cdots \neq X_n) = \prod_{i=0}^{n-1} \frac{k-i}{k} > \left(1 - \frac{n-1}{k}\right)^n \tag{S13}$$

We have not yet specified $k$; we do so now. Choose $k$ to be any integer with:

$$k \geq \frac{n-1}{1-q^{1/n}}$$

Then, combining this choice of $k$ with inequality S13, we have $P(X_1 \neq \cdots \neq Xn) > q$ substitute into inequality S12 to obtain $P(M_X) > q/n^n$, as required.

***Lemma 15.***

Let $X$ and $Y$ be independent random variables. Let $A$ be a set in the sample space of $X$ with $P(X \in A) > 0$. Let $B$ be a set in the product space of $X$ and $Y$. Assume that there exists a constant $p \in [0, 1]$ such that

$$P((x, Y) \in B) = p \text{ for all } x \in A$$

Then,

$$P((X, Y) \in B \mid X \in A) = p$$

**Proof:** Let $I_B(x, y)$ denote the indicator function for the event where $(x, y) \in B$. That is, $I_B(x, y) = 1$ if $(x, y) \in B$ and otherwise $I_B(x, y) = 0$. Let $g(x)$ be a function given by

$$g(x) = E[I_B(x, Y)] = P((x, Y) \in B)$$

Note that it follows from the conditions of this lemma that $g(x) = p$ for all $x \in A$. Consider the joint probability

$$P((X, Y) \in B, X \in A) = E[I_B(X, Y) I_A(X)]$$

where $I_A(x)$ is the indicator function for $x \in A$. From the law of total expectation, the above is equal to

$$= E[E(I_B(X, Y) I_A(X) \mid X)]$$

since $I_A(X)$ is a function of $X$ only, we may pull it out of the inner conditional expectation ("taking out what is known"; see for instance Theorem 9.3.2 in [45]):

$$= E[I_A(X) E(I_B(X, Y) \mid X)] \tag{S14}$$

Since $X$ and $Y$ are independent, it follows from Example 4.1.7 in [44] that $g(X) = E(I_B(X, Y) \mid X)$.

Substituting this result into Eq. S14 gives

$$= E[I_A(X) g(X)]$$

Notice that $I_A(x)g(x) = I_A(x)p$ because both evaluate to $p$ for all $x \in A$, and evaluate to 0 for all $x \notin A$. Thus the above becomes

$$= pE[I_A(X)]$$

To summarize this chain of equalities, we now have

$$P((X, Y) \in B, X \in A) = pE[I_A(X)] = pP(X \in A)$$

Finally, by substituting this result into the definition of conditional probability, we have

$$P((X, Y) \in B \mid X \in A) = \frac{P((X, Y) \in B, X \in A)}{P(X \in A)} = \frac{pP(X \in A)}{P(X \in A)} = p$$

as required.

## 3 The sequential permute-match test versus a seemingly similar flawed procedure

In the main text we briefly noted that the sequential permute-match test superficially resembles the flawed practice of conducting additional tests merely because earlier tests did not result in a detection. Here we provide a detailed explanation of why this practice generally results in an invalid *p*-value, and point out that the sequential permute-match test does not have this problem.

To spell out the flawed practice, suppose a researcher runs one test (test *X*) and obtains a *p*-value, $p_X$. If $p_X$ is below their desired significance level $\alpha$, they stop there and report $p_X$. Conversely, if $p_X > \alpha$, this researcher performs another test (test *Y*) that assesses the same biological hypothesis in a slightly different way, obtaining a second *p*-value, $p_Y$. However, the researcher has a vague understanding that if one conducts multiple tests, a correction is required, so at this point they perform a Bonferroni correction to account for two tests and report $2p_Y$. Overall, the *p*-value obtained from this series of steps is:

$$p = \begin{cases} p_X & \text{if } p_X \leq \alpha \\ 2p_Y & \text{otherwise} \end{cases} \tag{S15}$$

One may ask whether this overall procedure has the desired property (known as 'validity') that it prevents the false positive rate from exceeding the significance level. Unfortunately, this procedure is generally not valid, meaning that it commonly leads to $P(p \leq \alpha) > \alpha$ under the null hypothesis.

To demonstrate this fact mathematically, we make three assumptions. First, we assume that the null hypothesis of independence is true for both tests *X* and *Y* so that $P(p \leq \alpha)$ is the false positive rate. Second, we assume that $p_X$ and $p_Y$ are both random variables following a uniform distribution between 0 and 1. Finally, we assume that the significance level $\alpha$ is less than 1/2.

The false positive rate of the procedure in Eq S15 depends on the relationship between the two tests. If tests *X* and *Y* tend to report the same result, they will behave more like a single test and the false positive rate of Eq S15 will be lower; if the tests frequently report different results, they will behave like two genuinely distinct tests, and the false positive rate will be greater. This relationship is quantified in Table S1, which describes the joint and marginal distributions of the *X* and *Y* test outcomes. We write the probability that test *X* is significant (or not) along the bottom row, and the probability that test *Y* is significant (or not) along the rightmost column. We then define a parameter (*P*') which is the probability that test *Y* is significant while test *X* is not. We then fill out the rest of the table, corresponding to the probabilities of the other possible test results, by ensuring that the rows and columns of the 2 × 2 "Joint probability" table sum to the entries in the "Total probability" row and column.

| | | Joint probability | | Total probability |
|---|---|---|---|---|
| | | Test $X$ is significant ($p_X \leq \alpha$) | Test $X$ not significant ($p_X > \alpha$) | |
| Joint probability | Test $Y$ not significant ($2p_Y > \alpha$) | $\alpha/2 + P'$ | $1 - \alpha - P'$ | $1 - \alpha/2$ |
| | Test $Y$ is significant ($2p_Y \leq \alpha$) | $\alpha/2 - P'$ | $P'$ | $\alpha/2$ |
| Total probability | | $\alpha$ | $1 - \alpha$ | |

**Table S1. Joint and marginal probabilities of the outcomes of tests *X* and *Y*, assuming that $p_X$ and $p_Y$ both follow Unif(0, 1).** We first fill the total probability row and column, and define *P*' as the probability that test *Y* is significant but test *X* is not. The remaining entries can then be deduced.

Each joint probability is of course bounded between 0 and 1. Thus, moving clockwise from the upper left of the 2 × 2 inner table we have the following constraints on $P'$:

$$0 \leq \alpha/2 + P' \leq 1$$
$$0 \leq 1 - \alpha - P' \leq 1$$
$$0 \leq P' \leq 1$$
$$0 \leq \alpha/2 - P' \leq 1$$

Among these constraints, since $\alpha < 1/2$, the tightest bounds on $P'$ are $0 \leq P' \leq \alpha/2$. Notice that the false positive rate of the overall procedure is given by:

$$\begin{aligned} P(p \leq \alpha) &= 1 - P\left(p_X > \alpha \text{ and } 2p_Y > \alpha\right) \\ &= \alpha + P' \end{aligned} \tag{S16}$$

Fig S1 plots Eq S16 within the allowed bounds of $P'$, noting the extreme cases of $P' = 0$ and $P' = \alpha/2$, as well as the case of independence in which $P' = P\,(2p_Y \leq \alpha) \times (p_X > \alpha)$. Other than the extreme case of $P' = 0$, in which a significant $Y$ test result deterministically ensures that the $X$ test is also significant, the false positive rate of the overall procedure, $P\,(p \leq \alpha)$, will be greater than $\alpha$. Thus the procedure is not valid.

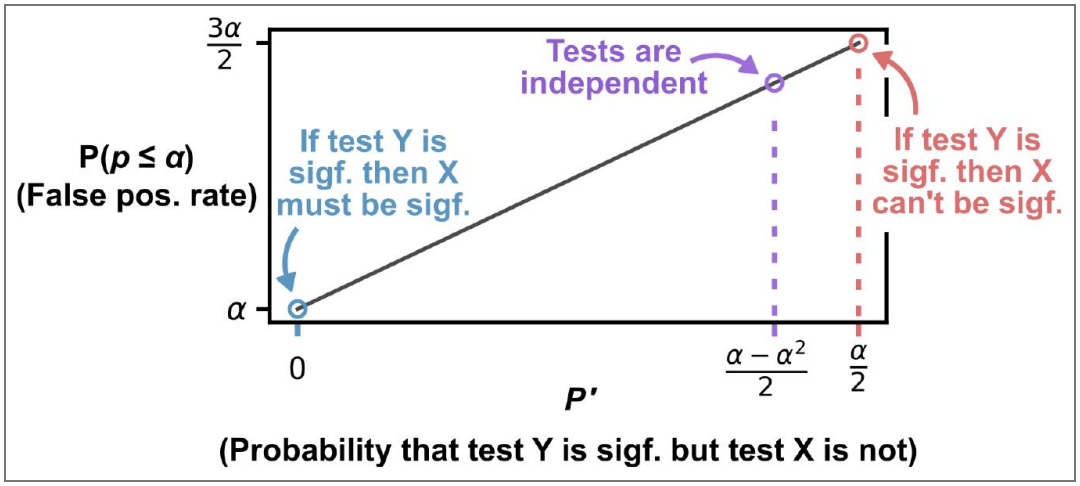


**Figure S1. The procedure of Eq S15 produces a false positive rate that generally exceeds *α*, and depends on the relationship between tests *X* and *Y*.** This relationship between *X* and *Y* can be quantified by *P'*, the probability that test *Y* is significant (i.e. $2p_Y \leq \alpha$) but test *X* is not significant (i.e. $p_X > \alpha$). The false positive rate is shown over the full range of *P'*, from 0 to *α*/2. The only way for the overall procedure of Eq S15 to be valid is when $P' = 0$, where tests *X* and *Y* are so tightly coupled that a significant result in test *Y* deterministically ensures that test *X* is also significant. Other than this edge case, the false positive rate of the overall procedure will exceed *α*.

The approach for the sequential permute-match test (Fig 1F) is different because it is based on discrete events, not continuous random variables. Perhaps counterintuitively, we can actually get away with a similar recipe while producing a valid *p*-value. To best compare our approach to the flawed procedure above, we presently leave out the permutation part and focus only on the matching test component in the following proposition:

> ***Proposition 16.***
>
> Let $M_X$ and $M_Y$ be events with $P\,(M_X) \leq 1/n^n$ and $P\,(M_Y) \leq 1/n^n$. Let

$$p = \begin{cases} 1/n^n & \text{if } M_X \text{ occurs} \\ 2/n^n & \text{if } M_Y \text{ occurs, but not } M_X \\ 1 & \text{neither occurs} \end{cases} \tag{S17}$$

Then, $P(p \le \alpha) \le \alpha$ for $0 \le \alpha \le 1$ and for all integers $n \ge 2$.

**Proof:** We consider four cases, each corresponding to possible values of $\alpha$.

First, if $\alpha = 1$, then $P(p \le \alpha) \le \alpha$ trivially because probability does not exceed 1.

Second is the case $1 > \alpha \ge 2/n^n$. In this case, either $M_X$ or $M_Y$ is sufficient to achieve $p \le \alpha$. So we have:

$$\begin{aligned} P(p \le \alpha) &= P(M_X \text{ or } M_Y) \\ &\le P(M_X) + P(M_Y) \le 2/n^n \le \alpha \end{aligned}$$

as required.

Third, consider the case $2/n^n > \alpha \ge 1/n^n$. In this case, only $M_X$ will give us $p \le \alpha$. So we have

$$P(p \le \alpha) = P(M_X) \le 1/n^n \le \alpha$$

as required.

Finally consider the case $\alpha < 1/n^n$. In this case $p \le \alpha$ is impossible so $P(p \le \alpha) = 0 \le \alpha$, as required.

Since the claim holds under our (exhaustive) set of cases, the proof is complete.

The same argument holds if $1/n^n$ and $2/n^n$ are replaced by $q$ and $2q$ for an arbitrary $q$ between 0 and 1, but we use $1/n^n$ and $2/n^n$ to highlight the connection to the permute-match test. Our complete proof (Theorem 11) followed largely the same steps as shown here. Overall, we hope that readers of this section now have a deeper understanding of why the sequential permute-match test is valid, and an appreciation of why similar approaches are invalid in the more standard context where statistical tests are not simply binary event checks.

## 4 Illustration of permute-match test with logistic map system

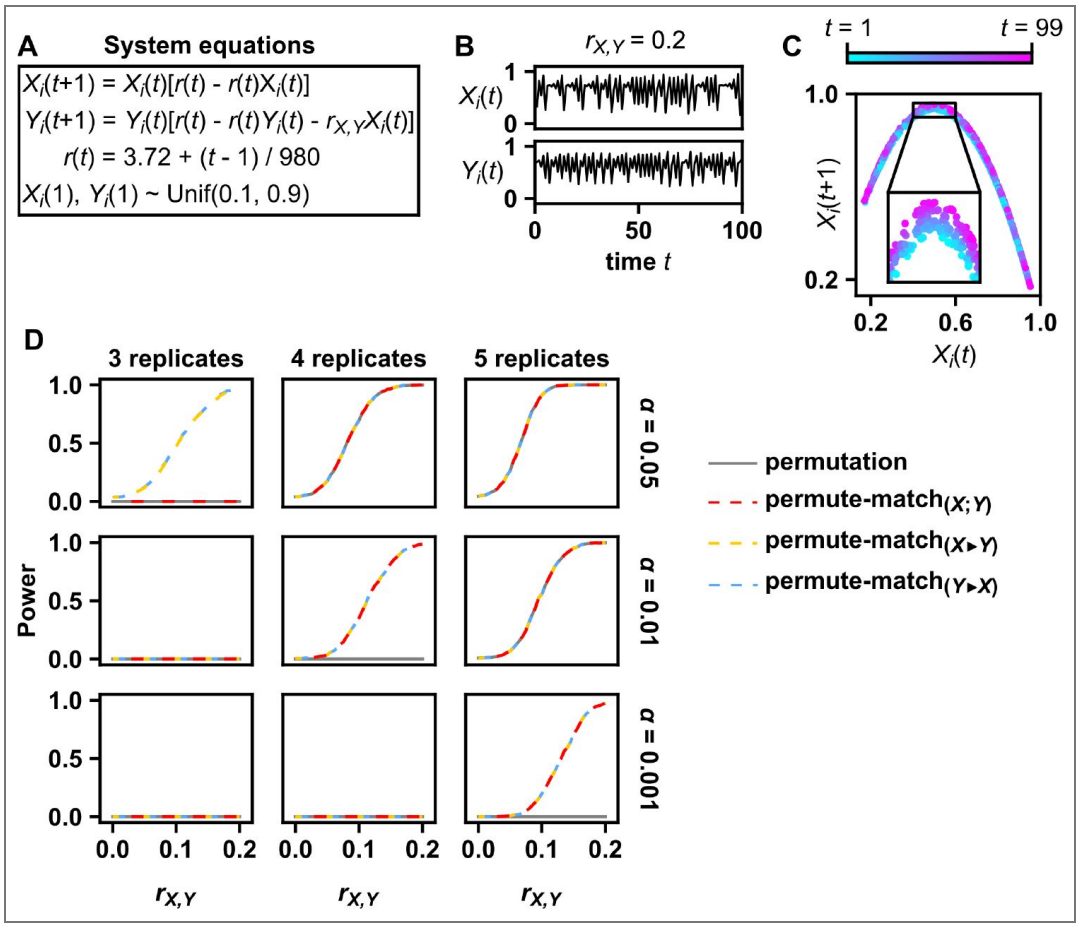


**Figure S2. Statistical power of permute-match tests and the permutation test in a nonlinear and nonstationary system.** (**A**) System equations. (**B**) Example dynamics. The processes $X$ and $Y$ are given by a coupled logistic map on a time grid $t = 1, 2, \ldots, 100$. The system is nonstationary because the parameter $r(t)$ varies with time from 3.72 when $t = 1$ to 3.82 when $t = 99$. (**C**) To see the nonstationarity clearly, we plot $X_i(t)$ against $X_i(t + 1)$ and color points by time, showing that the parabola that the points lie on is drifting over time. To better show the trend, 10 replicates are shown simultaneously in this chart. (**D**) Statistical power of the permutation test and permute-match tests as a function of the number of replicates, the significance level, and $r_{X,Y}$. Power was estimated from 5000 simulations at each value of $r_{X,Y}$ between $r_{X,Y} = 0$ and $r_{X,Y} = 0.2$ in steps of size 0.005. The correlation statistic ($\rho$) was cross-map skill, which is known to readily detect dependence between $X$ and $Y$ in this system [46]. For the cross-map skill calculation, we used $Y$ to estimate $X$, which corresponds to a scenario where the data analyst hypothesizes that $X$ influences $Y$. Cross-map skill requires two parameters, the embedding dimension and the embedding lag, and these were set to 2 and 1 respectively following prior works that used the logistic map for benchmarking [46, 47].

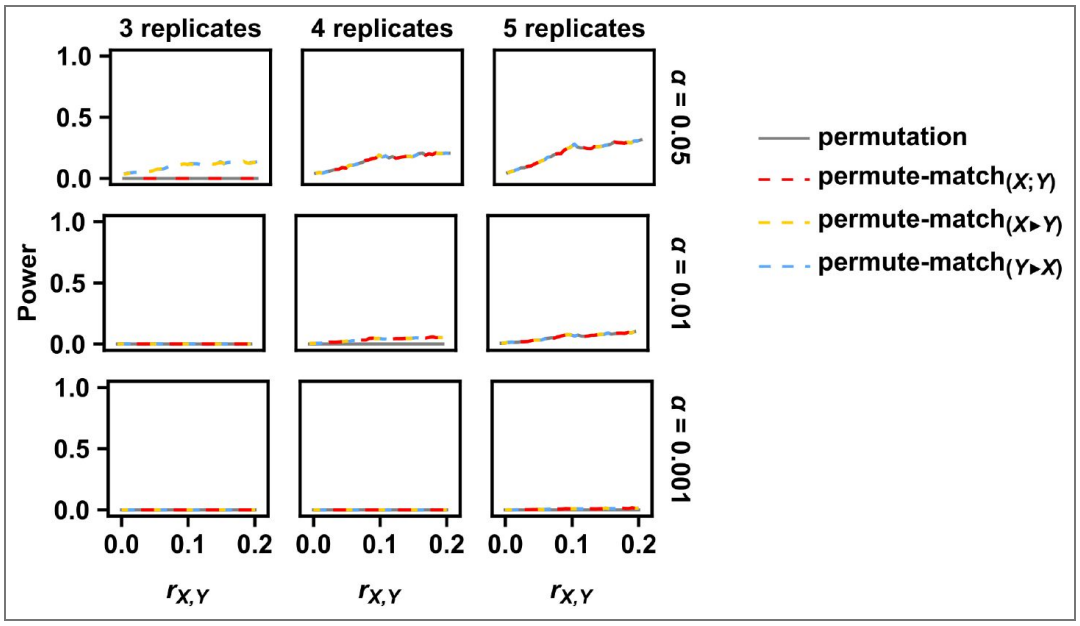


**Figure S3. Pearson correlation has lower power than cross-map skill for the logistic map.** Shown here is the same result as in Fig S2D, except now the correlation statistic ($\rho$) is Pearson correlation as in Fig 2.

# Additional information

## Funding


| Funder | Grant reference number | Author |
|---|---|---|
| HHS \| NIH \| National Institute of General Medical Sciences (NIGMS) | R01GM124128 | Alex E Yuan<br>Wenying Shou |
| Academy of Medical Sciences (The Academy of Medical Sciences) | | Alex E Yuan<br>Wenying Shou |
| National Science Foundation (NSF) | 1917258 | Wenying Shou |
| Royal Society (The Royal Society) | | Wenying Shou |
| Fred Hutchinson Cancer Center (FHCRC) | | Alex E Yuan |


## Author ORCID iDs

**Alex E Yuan**: https://orcid.org/0000-0002-8972-7497

# References


[1] **Granger CW,** Newbold P (1974) Spurious regressions in econometrics. *Journal of econometrics* **2**:111-20 https://doi.org/10.1016/0304-4076(74)90034-7

[2] **Yule GU** (1926) Why do we sometimes get nonsense-correlations between Time-Series?–a study in sampling and the nature of time-series. *Journal of the royal statistical society* **89**:1-63 https://doi.org/10.2307/2341482

[3] **Weiss S,** Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal* **10**:1669-81 https://doi.org/10.1038/ismej.2015.235 | PubMed

[4] **Coenen AR,** Weitz JS (2018) Limitations of correlation-based inference in complex virus-microbe communities. *MSystems* **3**:e00084-18 https://doi.org/10.1128/msystems.00084-18 | PubMed

[5] **Rosenthal JS** (2006) *First Look At Rigorous Probability Theory, A* World Scientific Publishing Company.

[6] **Ross SM** (2014) *A first course in probability* Pearson.

[7] **Peters J,** Janzing D, Schölkopf B (2017) *Elements of causal inference: foundations and learning algorithms* MIT press.

[8] **Hitchcock C,** Rédei M (2020) Reichenbach's Common Cause Principle. In: Zalta EN (Ed). *The Stanford Encyclopedia of Philosophy* Metaphysics Research Lab, Stanford University.

[9] **Chan KH,** Hayya JC, Ord JK (1977) A note on trend removal methods: The case of polynomial regression versus variate differencing. *Econometrica: Journal of the Econometric Society* 737-44 https://doi.org/10.2307/1911686

[10] **Isliker H,** Kurths J (1993) A test for stationarity: finding parts in time series apt for correlation dimension estimates. *International Journal of Bifurcation and Chaos* **3**:1573-9 https://doi.org/10.1142/S0218127493001227

[11] **Guarin D,** Orozco A, Delgado E (2010) A new surrogate data method for nonstationary time series. *arXiv* https://doi.org/10.48550/arxiv.1008.1804

[12] **Greene WH** (2012) *Econometric Analysis* Pearson.

[13] **Iatsenko D,** Bernjak A, Stankovski T, Shiogai Y, Owen-Lynch PJ, Clarkson P, et al. (2013) Evolution of cardiorespiratory interactions with age. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**:20110622 https://doi.org/10.1098/rsta.2011.0622 | PubMed

[14] **Lancaster G,** Iatsenko D, Pidde A, Ticcinelli V, Stefanovska A (2018) Surrogate data for hypothesis testing of physical systems. *Physics Reports* **748**:1-60 https://doi.org/10.1016/j.physrep.2018.06.001

[15] **Moulder RG,** Boker SM, Ramseyer F, Tschacher W (2018) Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses. *Psycho-logical methods* **23**:757 https://doi.org/10.1037/met0000172 | PubMed

[16] **Schreiber T,** Schmitz A (2000) Surrogate time series. *Physica D: Nonlinear Phenomena* **142**:346-82 https://doi.org/10.1016/s0167-2789(00)00043-9

[17] **Yuan AE,** Shou W (2022) Data-driven causal analysis of observational biological time series. *eLife* **11**:e72518 https://doi.org/10.7554/eLife.72518 | PubMed

[18] **Martin DC,** Buffington V, Becker J (1981) Randomization test of paired data: Application to evoked responses. *Psychophysiology* **18**:524-8 https://doi.org/10.1111/j.1469-8986.1981.tb01821.x | PubMed

[19] **Ernst MD** (2004) Permutation methods: a basis for exact inference. *Statistical Science* 676-85 https://doi.org/10.1214/088342304000000396

[20] **Buffington V,** Martin DC, Becker J (1981) VER similarity between alcoholic probands and their first-degree relatives. *Psychophysiology* **18**:529-33 https://doi.org/10.1111/j.1469-8986.1981.tb01822.x | PubMed

[21] **Albert M,** Bouret Y, Fromont M, Reynaud-Bouret P (2016) Surrogate data methods based on a shuffling of the trials for synchrony detection: the centering issue. *Neural Computation* **28**:2352-92 https://doi.org/10.1162/neco_a_00839 | PubMed

[22] **Weinstein SM,** Vandekar SN, Adebimpe A, Tapera TM, Robert-Fitzgerald T, Gur RC, et al. (2021) A simple permutation-based test of intermodal correspondence. *Human brain mapping* **42**:5175-87 https://doi.org/10.1002/hbm.25577 | PubMed

[23] **Bacchetti P,** Deeks SG, McCune JM (2011) Breaking free of sample size dogma to perform innovative translational research. *Science translational medicine* **3**:87ps24-4 https://doi.org/10.1126/scitranslmed.3001628 | PubMed

[24] **Vesterinen HV,** Egan K, Deister A, Schlattmann P, Macleod MR, Dirnagl U (2011) Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the Journal of Cerebral Blood Flow and Metabolism. *Journal of Cerebral Blood Flow & Metabolism* **31**:1064-72 https://doi.org/10.1038/jcbfm.2010.217 | PubMed

[25] **Ristić-Djurović JL,** Ćirković S, Mladenović P, Romčević N, Trbovich AM (2018) Analysis of methods commonly used in biomedicine for treatment versus control comparison of very small samples. *Computer Methods and Programs in Biomedicine* **157**:153-62

https://doi.org/10.1016/j.cmpb.2018.01.026 | PubMed

[26] **Frasch MG,** Walter B, Herry CL, Bauer R (2021) Multimodal pathophysiological dataset of gradual cerebral ischemia in a cohort of juvenile pigs. *Scientific Data* **8**:4 https://doi.org/10.1038/s41597-020-00781-y | PubMed

[27] **Vega ML,** Willemoes M, Thomson RL, Tolvanen J, Rutila J, Samaš P, et al. (2016) First-time migration in juvenile common cuckoos documented by satellite tracking. *PLoS One* **11**:e0168940 https://doi.org/10.1371/journal.pone.0168940 | PubMed

[28] **Greene CS,** Garmire LX, Gilbert JA, Ritchie MD, Hunter LE (2017) Celebrating parasites. *Nature genetics* **49**:483-4 https://doi.org/10.1038/ng.3830 | PubMed

[29] **Hemerik J,** Goeman J (2018) Exact testing with random permutations. *Test* **27**:811-25 https://doi.org/10.1007/s11749-017-0571-1 | PubMed

[30] **Shaw E** (1978) Schooling fishes. *American Scientist* **66**:166-75

[31] **Miller N,** Gerlai R (2012) From schooling to shoaling: patterns of collective motion in zebrafish (Danio rerio). *PloS one* **7**:e48865 https://doi.org/10.1371/journal.pone.0048865 | PubMed

[32] **Viscido SV,** Parrish JK, Grünbaum D (2004) Individual behavior and emergent properties of fish schools: a comparison of observation and theory. *Marine Ecology Progress Series* **273**:239-49 https://doi.org/10.3354/meps273239

[33] **Lopez U,** Gautrais J, Couzin ID, Theraulaz G (2012) From behavioural analyses to models of collective motion in fish schools. *Interface focus* **2**:693-707 https://doi.org/10.1098/rsfs.2012.0033 | PubMed

[34] **Romero-Ferrero F,** Bergomi MG, Hinz RC, Heras FJ, De Polavieja GG (2019) Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. *Nature methods* **16**:179-82 https://doi.org/10.1038/s41592-018-0295-5 | PubMed

[35] **Berens P** (2009) CircStat: a MATLAB toolbox for circular statistics. *Journal of statistical software* **31**:1-21 https://doi.org/10.18637/jss.v031.i10

[36] **Virtanen P,** Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**:261-72 https://doi.org/10.17863/cam.73648

[37] **Kwiatkowski D,** Phillips PC, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of econometrics* **54**:159-78 https://doi.org/10.1016/0304-4076(92)90104-y

[38] **Kulkarni RU,** Wang CL, Bertozzi CR (2022) Analyzing nested experimental designs-A user-friendly resampling method to determine experimental significance. *PLoS computational biology* **18**:e1010061 https://doi.org/10.1371/journal.pcbi.1010061 | PubMed

[39] **Cannistra C,** Hoang L, Yuan AE, Shou W (2024) Subtle methodological variations substantially impact correlation test results in ecological time series. *bioRxiv* https://doi.org/10.1101/2024.10.11.617506

[40] **Yuan AE,** Shou W (2022) An exactly valid and distribution-free statistical significance test for correlations between time series. *bioRxiv* https://doi.org/10.1101/2022.01.25.477698

[41] **Seabold S,** Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. **57** Austin, TX. pp. 10-25080 https://doi.org/10.25080/majora-92bf1922-011

[42] **Lehmann EL,** Romano JP (2005) *Testing statistical hypotheses* **3** Springer.

[43] **Steele JM** (2004) *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities* Cambridge University Press.

[44] **Durrett R** (2019) *Probability: theory and examples* **49** Cambridge university press.

[45] **Blitzstein JK,** Hwang J (2019) *Introduction to probability* Chapman and Hall/CRC.

[46] **Sugihara G,** May R, Ye H, Ch Hsieh, Deyle E, Fogarty M, et al. (2012) Detecting causality in complex ecosystems. *Science* **338**:496-500 https://doi.org/10.1126/science.1227079 | PubMed

[47] **Ye H,** Deyle ER, Gilarranz LJ, Sugihara G (2015) Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports* **5**:14750 https://doi.org/10.1038/srep14750 | PubMed

**Romero-Ferrero F,** Bergomi MG, Hinz RC, Heras FJ, De Polavieja GG (2019) Idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. Google Drive. ID 1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD https://drive.google.com/drive/folders/1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD

# Peer reviews

## Reviewer #1 (Public review):

[Editors' note: this version has been assessed by the Reviewing Editor without further input from the original reviewers. The authors have addressed the comments raised in the previous round of review: Definitions and terminology have been made more precise. Additional analysis confirms the conclusions previously stated and clarifies concerns about the computational tractability of the method.]

Summary:

The manuscript puts forward a statistical method to more accurately report the significance of correlations within data. The motivation for this study is two-fold. First, the publication of biological studies demands the report of p-values, and it is widely accepted that p-values below the arbitrary threshold of 0.05 give the authors of such studies justification to draw conclusions about their data. Second, many biological studies are limited by the number of replicate samples that are feasible, with replicates of less than 5 typical. The authors report a statistical tool that uses a permute-match approach to calculate p-values. Notably, the proposed method reduces p-values from around 0.2 to 0.04 as compared to a standard permutation test with a small sample size. The approach is clearly explained, including detailed mathematical explanations and derivations. The advantage of the approach is also demonstrated through analysis of computer-generated synthetic data with specified correlation and analysis of previously published data related to fish schooling. The authors make a clear case that this method is an improvement over the more standard approach currently used and also demonstrate the impact of this methodology on the ability to obtain p-values that are the standard for biological research. Overall, this paper is very strong. While the subject matter seems somewhat specialized, I would make the case that this will be an important study that has broad general interest to readers. The findings are very general and applicable to many research contexts. Experimentalists also want to report accurate p-values in their work and better understand how these values are calculated. Although I believe the previous statement is true, I am not sure that many research groups doing biological work are reading specialized statistics journals regularly. Therefore, a useful and broadly applicable statistical tool is well placed in this journal.

Strengths:

The proposed method is broadly applicable to many realistic datasets in many experimental contexts.

The power of this method was demonstrated with both real experimental data and "synthetic" data. The advantages of the tool are clearly reported. The zebrafish data is a great example dataset.

The method solves a real-life problem that is frequently encountered by many experimental groups in the biological sciences.

The writing of the paper is surprisingly clear, given the technical nature of the subject matter. I would not at all consider myself a statistician or mathematician, but I found the text easy to

follow. The authors did an impressive job guiding the reader through material that would often be difficult to grasp. The introduction was also well-written and clearly motivated the goals of the study.

https://doi.org/10.7554/eLife.103703.2.sa2

## Reviewer #2 (Public review):

Summary:

This paper presented a hypothesis testing procedure for the independence of two time-series that was potentially suitable for nonlinear dependence and for small-sample cases. This should bring potential benefits for biology data.

Strengths:

The test offers good flexibility for different kinds of dependence (through adjusting \rho) and seems to have good finite sample performance compared to the literature. The justification regarding the validity of the test procedure is clear.

https://doi.org/10.7554/eLife.103703.2.sa1

## Author response:

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *The manuscript puts forward a statistical method to more accurately report the significance of correlations within data. The motivation for this study is two-fold. First, the publication of biological studies demands the report of p-values, and it is widely accepted that p-values below the arbitrary threshold of 0.05 give the authors of such studies justification to draw conclusions about their data. Second, many biological studies are limited by the number of replicate samples that are feasible, with replicates of less than 5 typical. The authors report a statistical tool that uses a permute-match approach to calculate p-values. Notably, the proposed method reduces p-values from around 0.2 to 0.04 as compared to a standard permutation test with a small sample size. The approach is clearly explained, including detailed mathematical explanations and derivations. The advantage of the approach is also demonstrated through analysis of computer-generated synthetic data with specified correlation and analysis of previously published data related to fish schooling. The authors make a clear case that this method is an improvement over the more standard approach currently used, and also demonstrate the impact of this methodology on the ability to obtain p-values that are the standard for biological research. Overall, this paper is very strong. While the subject matter seems somewhat specialized, I would make the case that this will be an important study that has broad general interest to readers. The findings are very general and applicable to many research contexts. Experimentalists also want to report accurate p-values in their work and better understand how these values are calculated. Although I believe the previous statement is true, I am not sure that many research groups doing biological work are reading specialized statistics journals regularly. Therefore a useful and broadly applicable statistical tool is well placed in this journal.*
>
> *Strengths:*

> *The proposed method is broadly applicable to many realistic datasets in many experimental contexts.*
>
> *The power of this method was demonstrated with both real experimental data and "synthetic" data. The advantages of the tool are clearly reported. The zebrafish data is a great example dataset.*
>
> *The method solves a real-life problem that is frequently encountered by many experimental groups in the biological sciences.*
>
> *The writing of the paper is surprisingly clear, given the technical nature of the subject matter. I would not at all consider myself a statistician or mathematician, but I found the text easy to follow. The authors did an impressive job guiding the reader through material that would often be difficult to grasp. The introduction was also well-written and clearly motivated the goals of the study.*

We appreciate the reviewer's summary of our study and its strengths.

> *Weaknesses:*
>
> *A few changes could be made if the manuscript is revised. I would consider all of these points minor, but the paper could be improved if these points were addressed.*
>
> *(1) The caption of Figure 2 doesn't seem to mention panel D. Figure A-2 also does not mention C in the caption.*

We apologize for this error, and thank you for catching it! The figure legends had missing or incorrect panel labels. This error has been corrected.

> *(2) Figure 2D is a little hard to follow. First, the definition of "Power" is not clear, and I couldn't find the precise definition in the text. Second, the legend for the different lines in 2D is only given in Figure A-2. Perhaps a portion of the caption for Figure 2 is missing?*

We have added a definition of power in the main text:

"Although the permutation test, simultaneous permute-match test, and sequential permute-match test are all valid, they vary in power – the probability of detecting true dependence."

We have clarified the use of "power" in legend of Fig 2 and clarified that the color key for Fig 2D is in Fig 2A. The relevant excerpt of the Fig 2 legend is copied here:

"(D) Statistical power for the permutation test and various permute-match tests as a function of the replicate number $n$, significance level $\alpha$, and strength of dependence $r_{X, Y}$. Power was estimated as the proportion of simulations in which dependence was detected, calculated from 5000 simulations at each value of $r_{X, Y}$ between $r_{X, Y} = 0$ and 0.54 in steps of size 0.01. At $r_{X, Y} = 0$, there is no dependence, so the curve at that point indicates the false positive rate rather than power. We chose the Pearson correlation coefficient as our correlation function $\rho$. See (A) for the color legend."

We have also added dotted lines connecting the legend in panel A to the curves in panel D.

> *(3) The concept of circular variance for the fish data was heard to understand/visualize. The equation on line 326 did not help much. If there is a very simple picture that could be added near line 326 that helps to explain Ct and theta, that could be a big help for some readers who do not work on related systems. The analysis performed is understandable, the reader just has to accept that circular variance captions the degree of alignment of the fish.*

We have replaced references to circular concentration with “mean resultant length”, which is the standard jargon for this term in circular statistics, and we have added an illustration.

> *(4) For the data discussed in Figure 3, I wasn't 100% sure how the time windows were selected. In the caption, it says "time series to different lengths starting from the first frame". So the 20 s time window was from t=0 to t= 20 s. Would a different result be obtained if a different 20 s window was chosen (from t = 4 min to t = 4 min 20 s just to give a specific example). I suppose by chance one of the time windows would give a pvalue less than the target 0.05, that wouldn't be surprising. Maybe a random time window should be selected (although I am not indicating what was reported was incorrect)? A little more discussion on this aspect of the study may be helpful.*

As suggested by the reviewer, we have redone the analysis of Figure 3D with random segments. This provides a more complete picture of how the chance of detecting a significant correlation varies with segment length. The main conclusion is unchanged: Perfect match tests reliably detect dependence across a wider range of segment lengths than the naive parametric alternative.

The relevant panel and an excerpt from the legend text are copied below.

“(D) Permute-match tests detected a significant correlation between speed and alignment more consistently than the parametric test. For a grid of lengths between 20 and 600 seconds we sampled 500 random segments of each length, each drawn from the first 600 seconds, and determined for each segment whether the parametric test and/or the two possible permute-match tests detected a significant ($p \leq 0.05$) correlation. In the edge case of the maximum 600-second length, all 500 “random” segments were identical.”

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *This paper presented a hypothesis testing procedure for the independence of two timeseries that was potentially suitable for nonlinear dependence and for small-sample cases. This should bring potential benefits for biology data.*
>
> *Strengths:*
>
> *The test offers good flexibility for different kinds of dependence (through adjusting \rho), and seems to have good finite sample performance compared to the literature. The justification regarding the validity of the test procedure is clear.*

We appreciate the reviewer’s summary of key aspects of our manuscript.

> *Weaknesses:*
>
> *(1) The size of the test is not guaranteed to (asymptotically) equal \alpha, which may damage the power.*

We thank the reviewer for raising the issue of test size and power. We agree that a conservative test (one whose size can fall below alpha) may sacrifice power.

Our objective is distribution-free false-positive rate (FPR) control. That is, we wish to keep the FPR at or below alpha for every distribution of X and Y, because in our regime (nonstationary time series with few independent replicates) the scientist often cannot verify distributional assumptions. Inspired by the reviewer's comment, we now show (new Proposition 14) that the perfect match probability can be made arbitrarily close to $1/n!$. As a consequence, any reported perfect match p-value below $1/n!$ would break the distribution-free validity of the test.

A test that exploits distributional structure could likely access lower p-values; we have now explored how the empirical FPR of the permute-match test varies with the data-generating process (see our response to reviewer 2's recommendation 1 below).

> *(2) The computational time can be an issue for a moderately large sample size when calculating the X / Y-perfect match. It will be beneficial to include discussions on the implementations of the test.*

We agree this is an important consideration. We have added the following text to the Discussion:

"The test appears computationally tractable for relevant sample sizes: Our implementation of the permute match$(X > Y)$ procedure completed a single test of dependence in the setting of Fig 2 with an average runtime of 3 seconds when n = 10 on a 2023 14-inch MacBook Pro with an M2 Pro processor and 16 GB RAM (see Source data 1). For $n > 10$, a standard permutation test already can report a $p$-value below $3 \times 10^{-8}$ so the perfect match test is likely unnecessary for typical applications."

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *A few more minor notes/comments:*
>
> *As a personal preference, I like it when figures printed in grayscale retain their meaning (when possible). Just FYI, Figure 3D in grayscale is uninterpretable. Not saying a change is needed, just pointing it out.*

We changed Fig 3D to address a comment above, and think the new version better distinguishes between the permute-match and permutation test results in greyscale.

> *On line 304, is that a lower bound or an upper bound? Maybe the issue is the probability mentioned on line 305 is not clear.*

As this point is not the main focus of the investigation, we have rephrased it to make it less technical and eliminate the issue of which bound is in question.

"It seems likely that the tests could be further modified to report an even lower $p$-value when an $X$- and $Y$ -perfect match occur simultaneously, as in Fig 3B. However, we have not investigated further and this problem is left for future efforts."

> *I am not sure if this is a weakness, but the p-value changing depending on the choice of whether to apply the X-perfect match or Y-perfect match test first is fascinating. The authors did discuss this very issue at several points in the manuscript. It is slightly unsettling to me that there isn't an exact p-value for a given set of data. This one point gives me a new perspective on statistics.*
>
> *In full transparency, I don't believe I have to background to thoroughly review the appendix. I did read through it and did not notice any errors, but I couldn't confidently say there are not any small mathematical errors or any logical flaws in the proofs. Some*

> *sections were not easy to follow (my own shortcomings, the writing appeared sufficient for more of an expert to understand).*

We greatly appreciate the reviewer's time and effort tackling an appendix outside their comfort zone.

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) In the numerical experiment session, the authors should include the null situation, i.e., the performance of the test when X and Y are independent. This helps assess the size of the test.*

We have added a section on size to our results section, copied below:

"The permute-match test's false positive rate depends on the process tested. The permute-match test is conservative – meaning that its false positive rate can fall below the significance level – because both the permutation test and perfect match test are conservative. As discussed elsewhere [29], the permutation test is conservative when $\alpha$ is not one of its possible $p$-values and when ties may occur between the original correlation and shuffled correlations. Checking for a perfect match is similarly conservative. The actual probability of a false-alarm perfect match event can vary depending on the process being tested. To see this consider the permute-match$(X > Y)$ test in the setting where $n$ = 3, where $\alpha$ = 0.05, and where $r_{X,Y}$= 0 (independent $X$ and $Y$). Note that in this case, obtaining a $Y$ -perfect match (and thus $p = 1/n^n$) is necessary and sufficient to detect dependence since α is too low for detection by either the permutation test or the $p = 2/n^n$ leg of the permute-match test. In the linear system of Fig 2, we observed among 5000 simulations a detection rate of 0.0148, significantly below the upper bound of $1/3^3$ (Figure 2 - Source data 1; one-tailed exact binomial test, $p < 10^{-20}$). Conversely, in the nonlinear system of Fig S2, this same event ($Y$ -perfect match under $n$ = 3 and $r_{X,Y}$ = 0) occurs with a detection rate of 0.0328, not significantly 9 below the upper bound of $1/3^3$ (Figure S2 - Source data 1; one-tailed exact binomial test, $p$ = 0.059). Thus, depending on the underlying process studied, the actual chance of a perfect match happening under independence may be near or significantly below the theoretical upper bound."

> *(2) Some insights regarding the choice of rho should be provided. Especially, are there any examples that the classical test, such as the Pearson correlation or Granger causality test does not work?*

We have redone the example of Appendix 4 with Pearson correlation, showing that Pearson correlation has substantially lower power than cross-map skill in this case (compare figures S2 and S3).

> *(3) Line 34 - 35, page 2: Correlations and causality should be separately considered. This sentence talks more about causality rather than correlation.*

We appreciate the reviewer's perspective and agree that correlation and causality are distinct.

We feel that pointing out the issue of spurious correlations is helpful to orient our readers, especially those from a broad scientific audience. In the text, we define "correlation" as a descriptive statistic (rather than normalized covariance), and later distinguish it from "dependence", which has causal implications due to Reichenbach's common cause principle. We believe this distinction provides a useful backdrop for practitioners who use statistical methods but are perhaps new to thinking deeply about statistical dependence.

> *(4) Please add some discussions on the situation that X_i depends on Y_{i - j} for some j > 0, which is associated with the setting of Granger causality test.*

We have added the following to the discussion:

"No distributional assumptions are required, and the correlation function ρ can be completely arbitrary. For instance, $\rho$ could include a lag to detect delayed dependence, or even evaluate the correlation strength at several lags and report the strongest among them [39]."

https://doi.org/10.7554/eLife.103703.2.sa0