The Trouble with Triples: Examining the Impact of Measurement Error in Mediation Analysis

Madeleine S. Gastonguay[1], Gregory R. Keele[1] and Gary A. Churchill[1,*]

[1]The Jackson Laboratory, Bar Harbor, ME 04609, USA

*Corresponding author: 600 Main Street, Bar Harbor, ME 04609, USA. Email: gary.churchill@jax.org

Abstract

Mediation analysis is a class of statistical methods used to determine whether the effect of an exogenous variable (X) on a target variable (Y) is mediated through an intermediate variable (M). Mediation analysis relies on propagation of variation through the variables in a causal system to produce characteristic patterns of correlation in the observed data. However, data often include measurement error that does not propagate, and in many applications of mediation analysis the potential effects of measurement error are not accounted for. In this work, we evaluate the reliability of mediation analysis in the presence of measurement error. We define a measurement error model and show that mediation analysis can infer partial mediation, *i.e.*, X affects Y both directly and indirectly through M, even when there is no causal relationship between M and Y. This becomes more likely as sample size increases. When partial mediation is excluded from the set of possible models, the relative measurement errors of X, M, and Y play a large role in determining whether mediation analysis will infer the correct causal relationship. Using examples from genetic studies of molecular traits, we demonstrate how to assess the reliability of mediation analysis in the presence of measurement error and highlight common scenarios that can lead to incorrect inferences.

Keywords: Quantitative trait loci; QTL; Bayesian model selection; partial mediation; causal inference; Sobel test

Introduction

Mediation analysis is a causal inference technique used to infer the relationship among three variables. Causal effects between variables propagate variation at each step, producing characteristic patterns of correlation in the data, which mediation analysis seeks to detect. However, in reality, the measurement of each variable incurs error external to the causal variation, representing a form of model mis-specification that could potentially bias mediation inference (Richmond *et al.* 2016). Our aim is to assess the reliability of mediation analysis in the presence of measurement error and to identify scenarios where the resulting inference may be misleading.

We consider a three-variable system in which the effect of an exogenous variable (X) on a target (Y) may be wholly, partially, or not transmitted through a mediator (M). The causal system can be represented as a directed acyclic graph as in Figure 1. The edges in this graph represent causal effects that may or may not be present. For example, if the effect of X on Y is wholly mediated through M, the edge labeled c is absent, indicating that there is no direct causal effect of X on Y. Common approaches to mediation analysis include the causal steps (Baron and Kenny 1986) and Sobel's test (Sobel 1982), which can establish the presence of an indirect effect of X on Y through M, *i.e.*, edges labeled a and b are both present. Here, we apply Bayesian model selection to infer the most likely structure of the causal system (Crouse *et al.* 2022). While the main conclusions of our work are not dependent on the choice of inference method, the Bayesian model selection approach is likelihood-based, performs as well as the Sobel test with the added advantage of distinguishing complete from partial mediation, and can incorporate multi-state exogenous variables and covariates. However, as with other commonly used mediation analysis methods, Bayesian model selection does not account for measurement error.

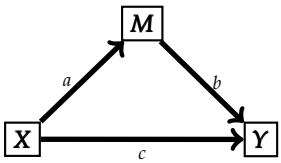

Figure 1 A simple mediation model. X effects Y directly (c) and indirectly through M (ab).

Previous studies have shown that measurement error in a mediator M results in underestimation of the indirect effect (the effect of X on Y through M) and loss of conditional independence between X and Y when the effect of X on Y is fully mediated through M, *i.e.*, complete mediation (Ledgerwood and Shrout 2011; Otter *et al.* 2018; Pierce *et al.* 2014; Rockman 2008). Furthermore, measurement error can lead to inference of partial mediation even when Y and M are independently affected by X (Otter *et al.* 2018; Pierce *et al.* 2014), a point we strongly emphasize here. Our work extends these observations beyond the context of complete mediation and considers the effects of measurement error in each of the measured variables.

We are particularly interested in the performance of mediation analysis in genetic studies where X is the genotype at a quantitative trait locus (QTL) associated with a target phenotype

Y, and the candidate mediator M is the transcript or protein expression level of a gene that co-localizes with the QTL, *i.e.*, a local gene expression or protein QTL (eQTL or pQTL). This is similar to transcript-wide association studies (TWAS) (Li and Ritchie 2021), an alternative causal inference approach to identifying gene expression mediators of traits. Co-localization of the target trait and the candidate mediator is not sufficient to establish a causal relationship because an association between M and Y could result from linkage disequilibrium (LD) between genetic variants with unrelated causal effects; it could be induced by an unobserved confounder of M and Y, a problem that instrumental variable (IV) causal inference methods, known as Mendelian randomization (MR) in genetic contexts, are designed to handle (Katan 1986; Didelez and Sheehan 2007); or, as we demonstrate below, an artifact due to measurement error.

Results

Measurement Error Models

To model measurement error in a three-variable causal system (Figure 1), we assume that the *causal variables* X^*, Y^*, and M^* are not directly observed. Instead we observe their surrogates, the *measured variables* X, Y, and M. We define the measurement error model in terms of six correlation parameters (Figure 2A; Appendix). The *causal correlations* ($\rho_{X^*M^*}$, $\rho_{X^*Y^*}$, and $\rho_{Y^*M^*}$) determine the relationships among the causal variables, which we refer to as the *causal structure* of the measurement error model. The *error correlations* (ρ_{X^*X}, ρ_{Y^*Y}, and ρ_{M^*M}) determine the level of measurement error between each causal variable and its measured counterpart. Low error correlations correspond to high measurement error, and vice-versa. We assume that measurement error is independent for each pair of variables. Adding features such as hidden confounders or correlated measurements would further obstruct correct inference of the causal structure.

The *data correlations* (ρ_{XY}, ρ_{XM}, and ρ_{YM}) that describe the expected correlations among the measured variables can be derived from the causal and error correlations:

$$\begin{aligned}\rho_{XY} &= \rho_{X^*X} \cdot \rho_{X^*Y^*} \cdot \rho_{Y^*Y}\\ \rho_{XM} &= \rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{M^*M}\\ \rho_{YM} &= \rho_{Y^*Y} \cdot \rho_{Y^*M^*} \cdot \rho_{M^*M}.\end{aligned}$$

The data correlations are always weaker than their corresponding causal correlations, more so when there is more measurement error. The three data correlations can be estimated from observed data, but we cannot estimate the causal and error correlations without additional information or constraints (see Appendix).

Our objective is to uncover the causal structure, *i.e.*, the relationships among the causal variables, given the measured surrogates. We adopt terminology from Schadt *et al.* (2005) and Neto *et al.* (2013) on genetic mediation analysis, and refer to the Causal, Independent, Reactive, and Complex models as the causal structures of interest (Figure 2B-E). In addition, there are causal structures for which the three causal variables are not fully connected that we refer to as non-mediation models. Under the Causal model, also known as complete mediation, the effects of X^* on Y^* are completely mediated through M^*. Under the Independent model, X^* has direct but independent effects on M^* and Y^*. Under the Reactive model, the roles of the mediator and target variables are reversed such that M^* is responding to variation in Y^*. Finally, under the Complex model, also known as partial mediation, X^* affects Y^* directly and indirectly through M^*. We aim to determine when it is possible to recover the correct causal structure from observed data with measurement error.

The Causal, Independent, and Reactive models impose constraints on the causal correlations. Specifically, the causal correlation between the two variables that are not directly connected by an edge is constrained to equal the product of the other two causal correlations. Equivalently, the partial correlation, *i.e.*, the association between two variables after removing the effect of a third, between the unconnected causal variables given the middle variable is equal to zero. In contrast, for the Complex model the causal correlations are not constrained, aside from the requirement that the correlation matrix is positive semidefinite.

Consider the Causal measurement error model with parameters as specified in Figure 2J. The causal correlations satisfy the constraint $\rho_{X^*Y^*} = \rho_{X^*M^*} \cdot \rho_{Y^*M^*}$ or equivalently, $\rho_{X^*Y^*|M^*} = 0$. However, the data correlations do not satisfy these constraints; even a modest amount of measurement error can result in data correlations that differ substantially from the causal correlations. The key contributor to this discrepancy for the Causal model is the error correlation for the mediator (ρ_{M^*M}), as the data correlations will satisfy the same constraints as the causal correlations if and only if there is no measurement error in the middle variable (see Appendix). Thus, if we ignore measurement error, the constraints imposed by the Causal, Independent, and Reactive models no longer hold; the data correlations will be unconstrained and thus will be consistent with the Complex model, *i.e.*partial mediation. Below, we examine whether this problem persists in the context of testing applied to data from finite samples.

The Latent Variable Model

The Causal, Independent, and Reactive measurement error models each have five free parameters (six parameters with one constraint) and three observable outcomes. However, each of the constrained models is equivalent to a latent variable model with only three free parameters. In the latent variable model, the causal variables X^*, M^*, and Y^* are replaced by a single *latent variable*, U (Figure 2F). The *latent correlations* (ρ_{XU}, ρ_{YU}, and ρ_{MU}) determine the relationship between U and each of the measured variables. For any of the constrained measurement error models, the data correlations can be expressed in terms of latent correlations,

$$\begin{aligned}\rho_{XY} &= \rho_{XU} \cdot \rho_{YU}\\ \rho_{XM} &= \rho_{XU} \cdot \rho_{MU}\\ \rho_{YM} &= \rho_{YU} \cdot \rho_{MU}.\end{aligned} \tag{1}$$

We can also express the latent correlations in terms of the causal and error correlations. In this case, the expressions depend on which of the constrained models is being considered. The latent correlations correspond to either the product of a causal correlation and an error correlation, or, for the middle variable, to an error correlation only (Table 1, Figure 2G-I). The latent variable model shows us that specific products of causal and error correlations can be uniquely identified, *e.g.*, $\rho_{Y^*Y} \cdot \rho_{Y^*M^*}$ for the Causal model. Importantly, for each constrained model, the error correlation of the middle variable is estimable (see Appendix).

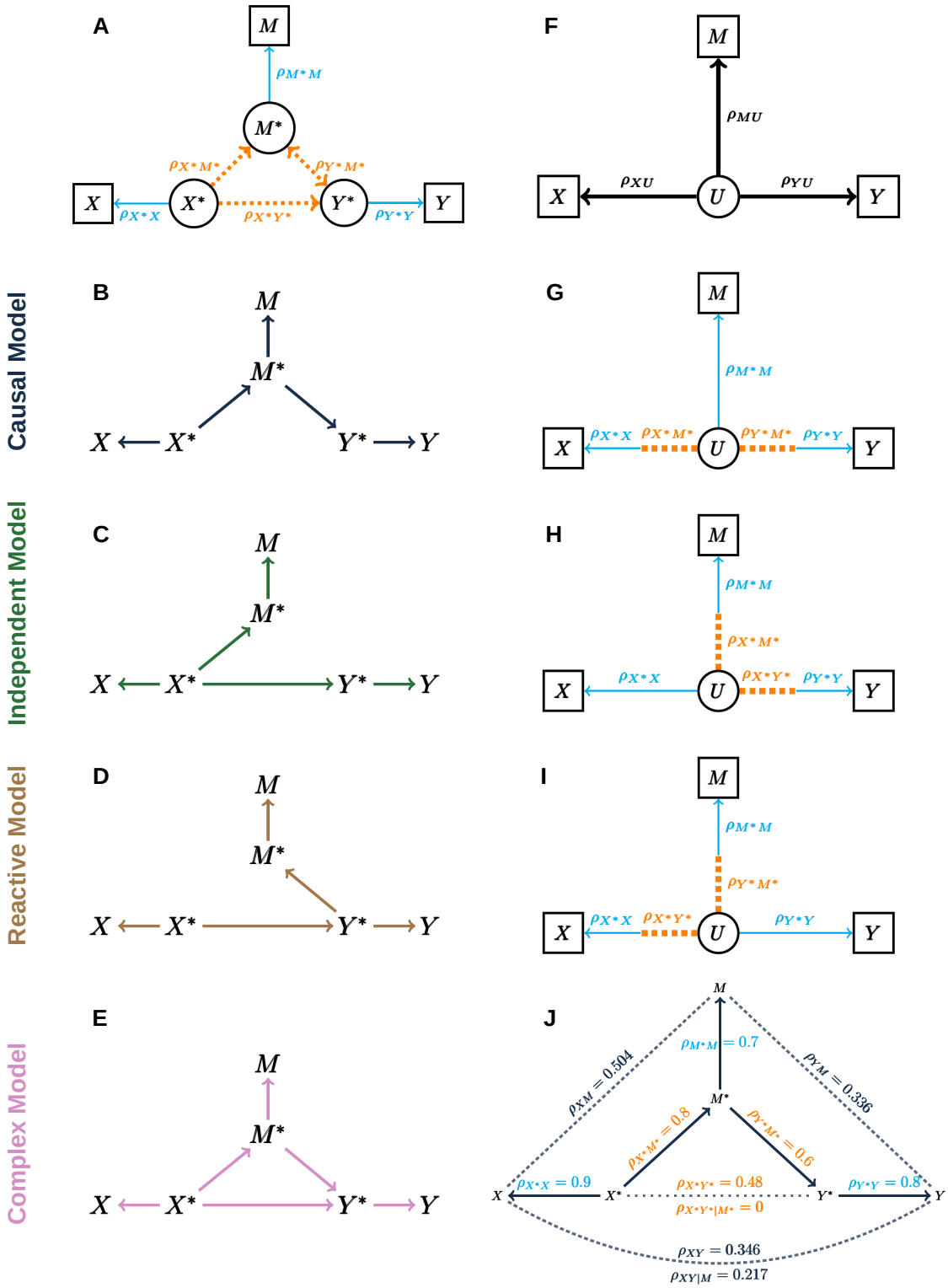

Figure 2 Directed acyclic graphs (DAGs) represent the structure of measurement error models and their corresponding latent variable models. (A) The measurement error model describes the relationships among causal variables X^*, Y^*, and M^* in terms of their causal correlations (dotted orange lines) and between each causal variable and its corresponding measured variable in terms of error correlations (solid blue lines). Variables in circles are unobserved and those in boxes are measured. X represents an exogenous variable, M represents a candidate mediator, and Y represents a target. (B-E) DAGs for the Causal, Independent, Reactive, and Complex measurement error models. (F) The structure of the latent variable model where U is a single unobserved (latent) variable. (G-I) The latent correlations are determined by the causal correlations (dotted orange edges) and error correlations (solid blue edges) in different combinations for the Causal, Independent, and Reactive models, respectively. (J) An example measurement error model for the Causal model. Causal (partial) correlations are labeled in orange, error correlations are labeled in bright blue, and data (partial) correlations are labeled in dark blue. Dotted lines denote correlations between variables that do not share a direct edge in the measurement error model.

Table 1 Latent Correlations for each model.

Causal Structure	Latent Correlation		
	ρ_{MU}	ρ_{XU}	ρ_{YU}
Causal	ρ_{M^*M}	$\rho_{X^*X} \cdot \rho_{X^*M^*}$	$\rho_{Y^*Y} \cdot \rho_{Y^*M^*}$
Independent	$\rho_{M^*M} \cdot \rho_{X^*M^*}$	ρ_{X^*X}	$\rho_{Y^*Y} \cdot \rho_{X^*Y^*}$
Reactive	$\rho_{M^*M} \cdot \rho_{Y^*M^*}$	$\rho_{X^*X} \cdot \rho_{X^*Y^*}$	ρ_{Y^*Y}

Simulations

The parameters of the measurement error model, denoted as ρ, can be thought of as correlations estimated from infinitely large data. We now consider what happens when model structure is inferred from correlations estimated from data from finite samples, denoted as r.

We simulated data from the measurement error model and then analyzed the data assuming that there is no measurement error. Specifically, we simulated X^* from a standard normal distribution and then simulated M^*, Y^*, X, M, and Y according to linear models with the desired causal or error correlations (Crouse *et al.* 2022) (see Methods). The sample size for simulated data ranged from $N = 200$ to $N = 5000$. We simulated 100,000 data sets for each of the Causal, Independent, and Reactive measurement error models. Model parameters (causal and error correlations) were sampled independently from beta distributions to obtain data correlations similar to those that arise in practice. We applied Bayesian model selection (Crouse *et al.* 2022) and selected the model with the greatest posterior probability for each simulated data set. We first restricted the model selection to choose among only the Causal, Independent, or Reactive models (Schadt *et al.* 2005; Neto *et al.* 2013) (three-choice model options). We then repeated the model selection including the Complex model and non-mediation models (expanded model options).

Our rationale for examining three-choice model selection is partly motivated by previous approaches to genetic mediation analysis, *e.g.*, Schadt *et al.* (2005). In addition, recognizing that the measurement error model is equivalent to partial mediation (Complex model), we expect that Bayesian model selection in large samples will infer the Complex model regardless of the underlying causal structure. Simulations and examples presented below demonstrate the utility of applying both expanded and three-choice model selection approaches in practice.

Three-choice model selection. The data correlations obtained from simulations of the Causal, Independent, and Reactive models had substantially overlapping ranges, demonstrating that we cannot determine the causal structure based solely on the observed data (Figure 3). Overall, we correctly classified the causal structure for $\sim 62\%$ of simulated data sets with $N = 200$ (Table 2). The rate of correct classification across all parameter configurations increased with increasing sample size, but never exceeded 65% for sample sizes up to $N = 5000$ (Figure S1).

Bayesian three-choice model selection always selected the model that has no direct effect between the two causal variables whose observed data correlation is the weakest (see shading in Figure 3). For example, if the correlation between X and Y is less (in magnitude) than the other two data correlations, the Causal model will be selected because it does not include a direct effect of X^* on Y^*. This simple inference rule for three-choice model selection only holds for univariate X, M, and Y.

Expanding model options reduces incorrect model selection. When we expanded the model selection options to include the Complex and non-mediation models, we saw a decrease in the overall rate of correct classification (41% when $N = 200$) but an even greater proportional decrease in the rate of incorrect classification as one of the constrained models (9% when $N = 200$, Table 2). When $N = 200$, the Complex model was selected almost as frequently as the correct causal structure and for larger sample sizes, the rate of selecting the Complex model increased, *e.g.*, up to 88% when $N = 5000$ (Figures S1 and S2).

When using the expanded model selection options, the simple inference rule (shading) in Figure 4 no longer applied, but there were some regularities. The Complex model was selected when all three data correlations were similar in magnitude. A non-mediation model was selected when at least two of the data correlations were sufficiently weak. However, if one of the three constrained models was selected, it aligned with the inference rule from three-choice model selection.

Latent correlations determine the consistency of mediation analysis. Our ability to consistently infer the correct causal structure in the presence of measurement error depends on the parameters of the underlying measurement error model. Specifically, the latent correlations (Table 1) determine the large sample behavior of model selection (Figures S3 and S4). For example, if ρ_{MU} is the strongest latent correlation, then ρ_{XY} will be the weakest data correlation (Equation 1) and according to our inference rule, three-choice model selection will tend to select the Causal model as sample size increases. Similarly, when ρ_{XU} or ρ_{YU} are the strongest latent correlations, the Independent and Reactive models will be selected more frequently as sample size increases, respectively. Thus, it is the relative sizes of the latent correlations that determine whether Bayesian model selection will be consistent (tending toward the correct causal structure) or inconsistent (tending toward an incorrect causal structure). We observed these patterns in simulations of the Causal model with increasing sample size (Figure 5A).

We also looked at the influence of latent correlations on Bayesian model selection with expanded model options (Figure 5B). We found that for simulations of the Causal model, a strong ρ_{MU} was required to correctly select the Causal model. The Complex model was frequently selected at weaker values of ρ_{MU}, regardless of the strength of ρ_{XU} and ρ_{YU}. This misclassification was more prevalent with larger sample sizes; even when there is very little error in the mediator ($\rho_{M^*M} = 0.95$), the Causal model was selected for only 25% of data sets simulated with $N = 5000$. The remaining data sets were classified as Complex. These results were symmetric for the Independent and Reactive models, making it clear that mediation analysis tends to infer partial mediation when there is any noise in the middle variable.

Using latent correlations to determine the reliability of mediation analysis. The latent variable model can be used to draw conclusions about the reliability of a mediation inference in the presence of un-modeled measurement error. Recall that the latent correlations are products of causal and error correlations. Thus, while we cannot uniquely identify the causal and error correlations, if we know the true causal structure, we can constrain the range of possible values that would lead to the selected model (Figures 6, S5, and S6).

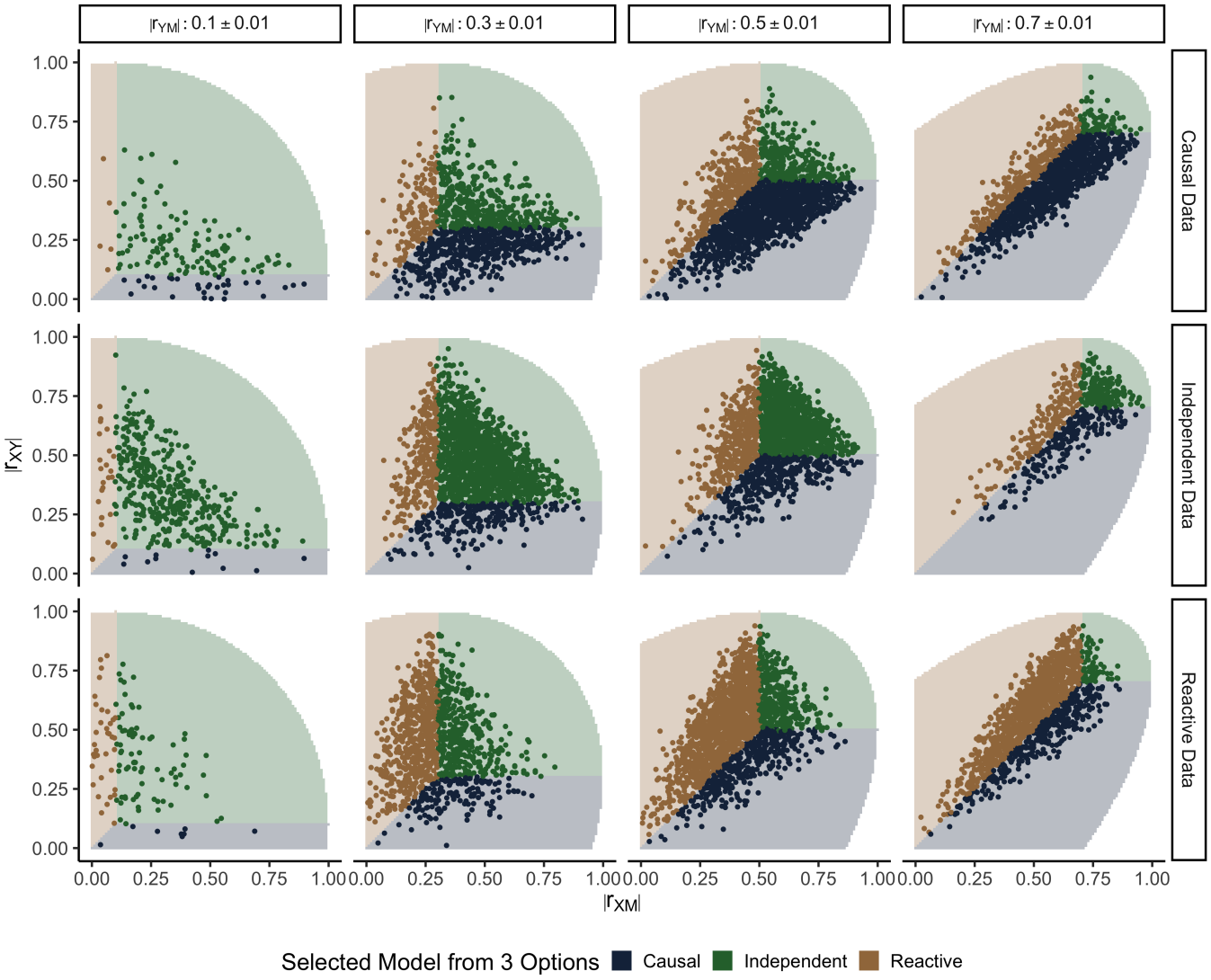

Figure 3 Estimated data correlations from simulations of each causal structure with $N = 200$ classified with three-choice model options. Each row of panels corresponds to simulations of a causal structure with $N = 200$. Columns correspond to binned values of r_{YM}. The x- and y-axes show r_{XM} and r_{XY}, respectively. Points representing the estimated data correlation from simulated data are colored to indicate the model with the greatest posterior probability from three-choice model selection. Shaded regions indicate the range of data correlations in which each model will be inferred, and the unshaded region delineates where the correlation matrices are not positive semi-definite. See Figure 4 for classifications based on expanded model options.

Table 2 Rates of accurate and inaccurate classification of simulated data across all parameter configurations when using three-choice and expanded model selection ($N = 200$).

Selected Model	Three-choice			Expanded		
	Causal (%)	Independent (%)	Reactive (%)	Causal (%)	Independent (%)	Reactive (%)
Causal	62.10	19.17	19.04	40.97	8.83	8.96
Independent	19.03	61.79	18.98	8.73	41.26	9.04
Reactive	18.87	18.98	61.98	8.87	8.95	41.20
Complex				38.63	38.68	37.97
Other				2.80	2.27	2.82

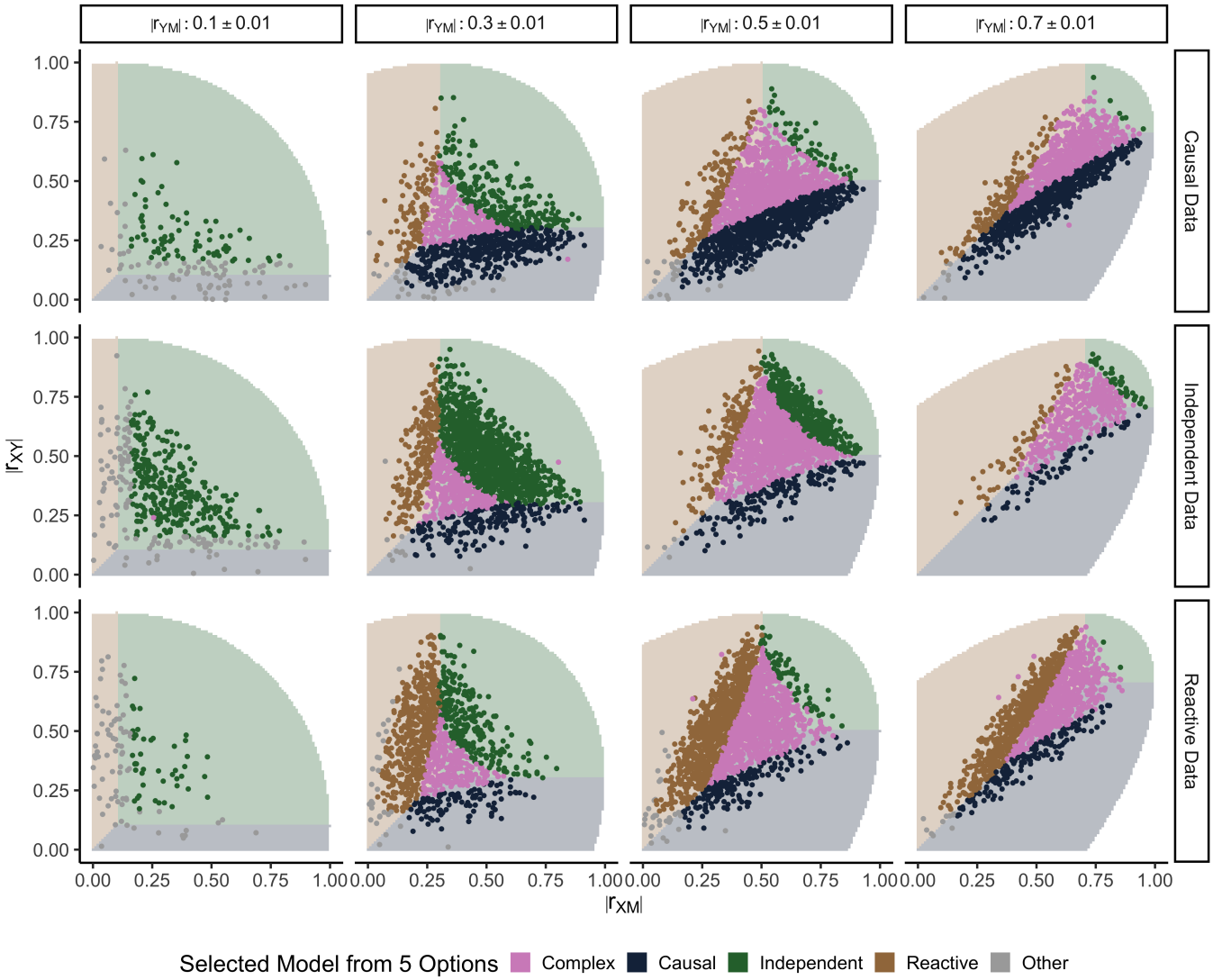

Figure 4 Estimated data correlations from simulations of each causal structure with $N = 200$ classified with expanded model options. Each row of panels corresponds to simulations of a causal structure with $N = 200$. Columns correspond to binned values of r_{YM}. The x- and y-axes show r_{XM} and r_{XY}, respectively. Points representing the estimated data correlation from simulated data are colored to indicate the model with the greatest posterior probability from expanded model selection. Shaded regions indicate the three-choice model selection inference rule, and the unshaded region delineates where the correlation matrices are not positive semi-definite. See Figure 3 for classifications based on three-choice model options and Figure S2 for results from simulations with $N = 5000$.

In practice, we will not know the true causal structure, but we can apply this reasoning in reverse. Suppose that the three-choice model selection infers the Causal model. If the Causal model is correct, the measurement error model parameters will be consistent with one of the DAGs in Figure 6A. There may be equal error in all three variables; there may be less error in the mediator than the other variables; or there may be more error in the mediator than the other variables, but the causal correlations are weak. Ironically, the latent variable model shows us that weaker causal correlations and/or more measurement error in the non-middle variables can improve our ability to select the correct casual structure. On the other hand, if the Causal model is selected but the true causal structure is Independent, there must be more error in the exogenous variable than in the candidate mediator and X^* should be tightly correlated with M^* (Figure S5B). Lastly, if the Causal model is selected and the true causal structure is Reactive, the measurement error for the target must be greater than for the mediator and Y^* is strongly correlated with M^* (Figure S6C). A complete enumeration of scenarios that lead to consistent or inconsistent inferences using three-choice model selection for each causal structure is outlined in Table S1.

Evaluating mediation analysis with real data

We obtained transcript and protein profiling data from liver tissue for 192 Diversity Outbred (DO) mice (Churchill *et al.* 2012), including mice of both sexes (Chick *et al.* 2016), and for 116 Collaborative Cross (CC) mice (Collaborative Cross Consortium 2012; Srivastava *et al.* 2017), with one female and one male from

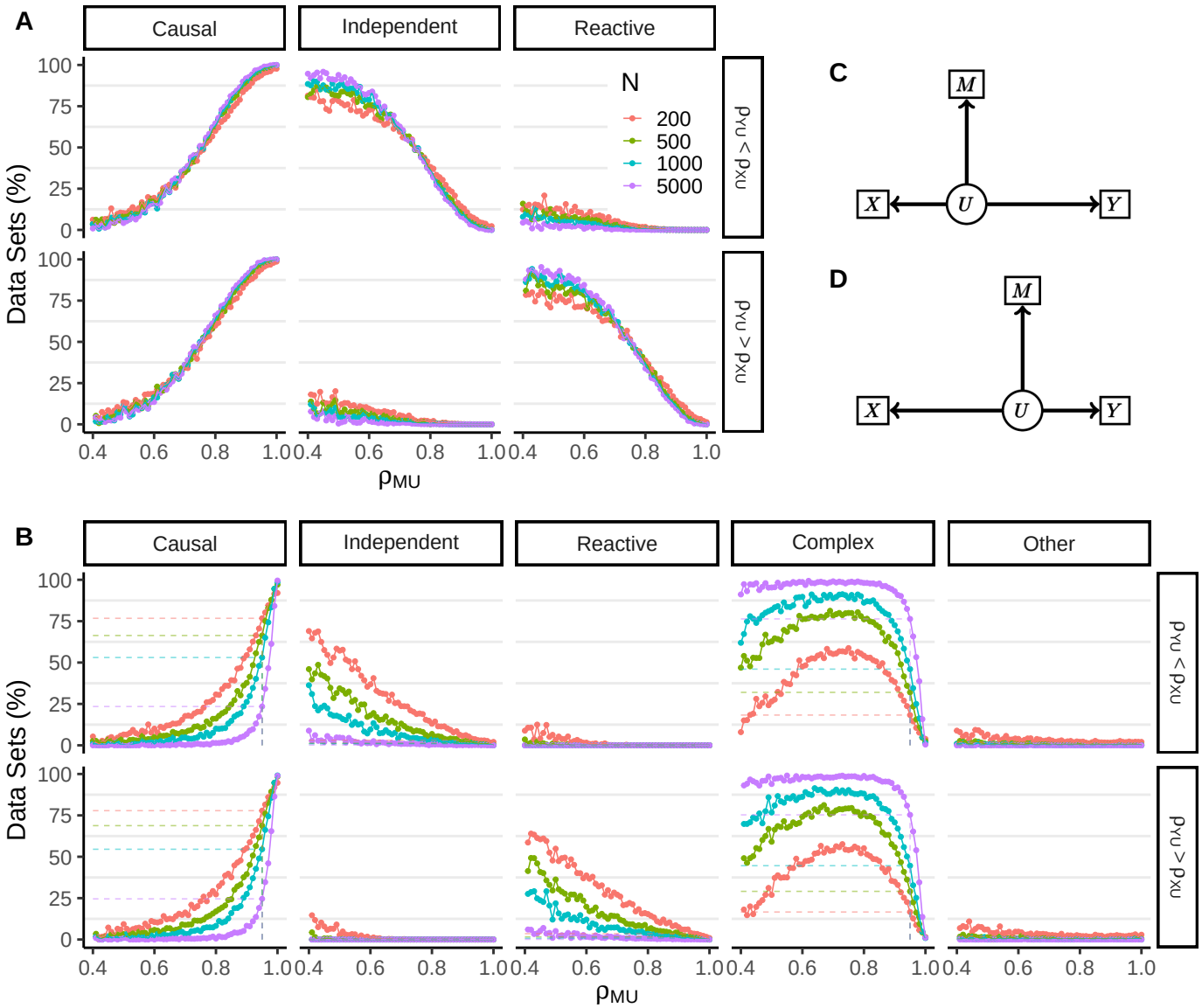

Figure 5 Rates of correct and incorrect classification as a function of error in the middle variable (*M*) for data simulated with the Causal model. Classification rates are shown for data simulated from the Causal measurement error model and classified with the three-choice (A) or expanded (B) model options. Line color denotes sample size used in simulations. Each column shows the percent of data sets classified as the model listed at the top as a function of the latent correlation ρ_{MU}. Dashed lines in (B) mark results when $\rho_{MU} = 0.95$. Both (A) and (B) are split into two rows showing results for data simulated with a stronger latent correlation for X (top row), and data sets with a stronger latent correlation for Y (bottom row). The corresponding latent variable DAGs are displayed in (C) and (D) where shorter edges denote stronger correlations.

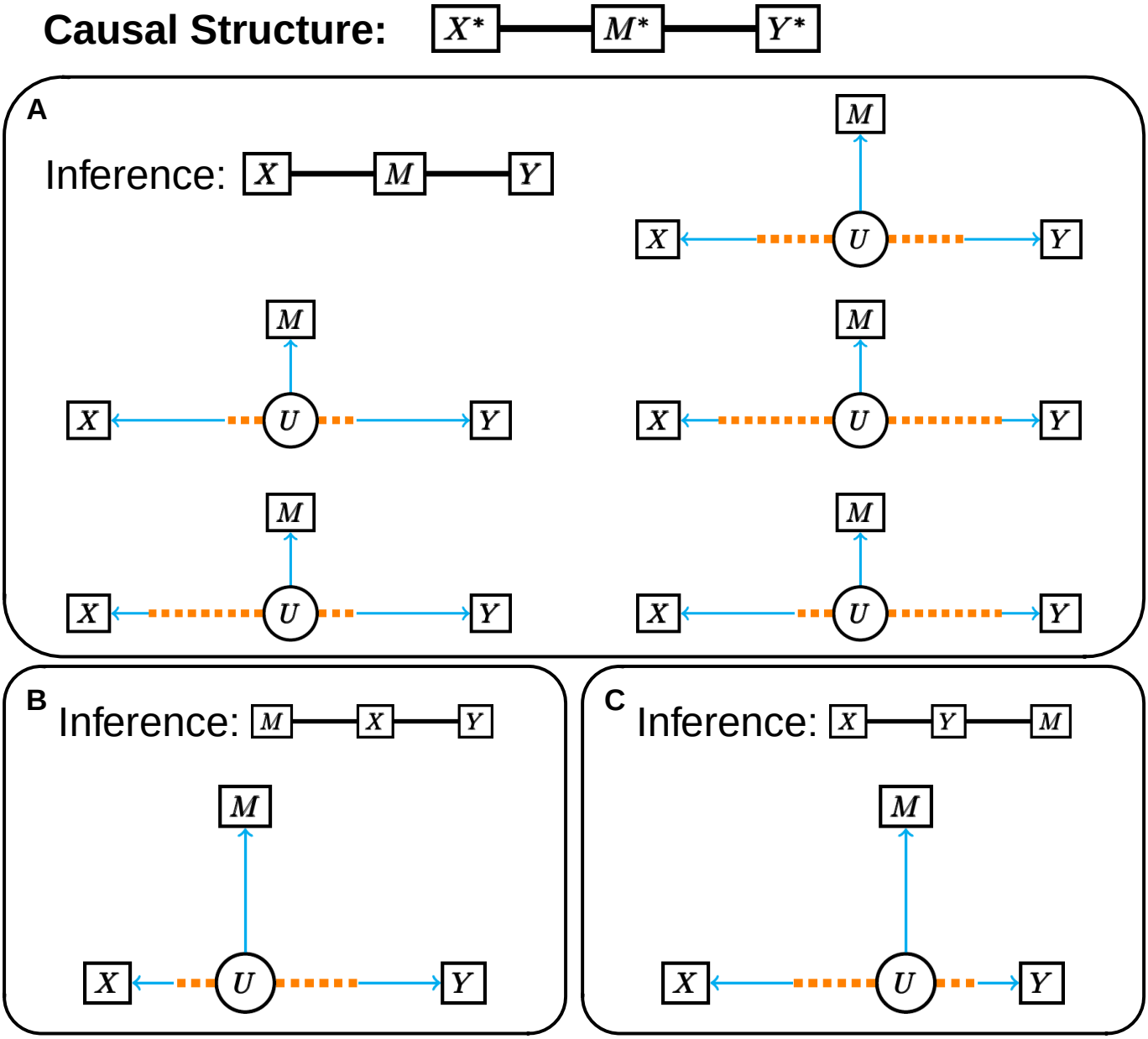

Figure 6 Latent variable representation of the causal measurement error model. Configurations of the latent variable model representing the Causal model that result in (A) consistent and (B-C) inconsistent inferences. Blue edges correspond to the proportion of the latent correlation determined by the error correlation and dotted orange edges correspond to the proportion determined by the causal correlation. Shorter edges represents a stronger correlation and vice-versa. (A) The correct model is inferred if the latent correlation for M is the strongest (the latent variable arm for M is the shortest). This can be achieved if there is an equal amount of error in all three variables (top right) or if there is less error in M than X and Y (middle left). If M is noisier than X and Y, the correct model may still be inferred if the causal correlations are weak (middle right). The bottom row shows scenarios where X and Y satisfy different configurations. (B) The Independent model is inferred if the latent correlation for X is the strongest. When the causal structure is the Causal model, this will only occur if the error correlation for M is weaker than both the error correlation and causal correlation contributing to the latent correlation for X. The composition of the latent correlation arm for Y does not influence the inference. (C) Shows the analogous scenario to (B) for inferring the Reactive model by swapping X and Y.

each of 58 CC strains (Keele *et al.* 2021). The DO and CC mice are descended from the same eight founder strains and thus share the same genetic variants (Saul *et al.* 2019). The DO mice are an outbred stock, of which each mouse is a genetically unique individual and predominantly heterozygous at loci across the genome. The CC mice represent a panel of recombinant inbred strains, homozygous at a vast majority of the genome, enabling the use of genetic replicates in experiments. The genomes of both DO and CC are mosaics of the eight founder strain haplotypes. To carry out genetic mediation analysis, we represented the genotype of each animal as an eight-state vector of haplotype dosages (Gatti *et al.* 2014), *i.e.*, X is a multi-state exogenous variable. The relationships between the multi-state genotype and the univariate transcript and protein abundances were defined by a regression model with eight regression coefficients that represent additive allele effects.

For each study, we identified genes with local protein abundance QTL (pQTL) and a corresponding local gene expression QTL (eQTL). We refer to these as concordant genotype-transcript-protein triplets and assume that in most cases the transcript mediates the effect of genetic variation on protein abundance. We found 2023 concordant triplets in the DO data, 967 in the CC data, and 582 genes with concordant triplets in both studies. For each concordant triplet, we identified a transcript from a nearby gene with the strongest co-mapping local eQTL within 1Mb of the pQTL. We refer to these triplets as discordant and assume that in most cases the genetic regulation of the transcript and protein occur independently. We applied three-choice and expanded model selection to each concordant and discordant triplet.

Allele effects. The 8-state exogenous variable provides more information than is available for univariate X (normal or binary), which we summarize as the estimated allele effects for M and Y, representing an 8-element vector. If the causal structure is Independent, we expect the correlation between the allele effects for M and Y to be randomly distributed, as was seen for the discordant triplets (Figure 7A). If the causal structure is Causal or Reactive, we expect the allele effects to be correlated, positively or negatively, as was seen for the concordant triplets (Figure 7B). The allele effects estimated across thousands of triplets support our assumption that the causal structure of most discordant triplets is Independent and that of most concordant triplets is Causal. However, even within a multi-allelic system, at times nearby genes may fall in the same LD region and have similar effects patterns. We observed that mis-classified discordant triplets tend to have a strong correlation between the RNA and protein allele effects. This problem may be a bigger issue in bi-allelic systems where the genetic effects represent less information. Thus, correlated allele effects cannot be used to discredit a particular causal structure.

Discordant triplets. Applying three-choice model selection to the discordant triplets, we found that 99% were classified as Independent in the CC and 98% were classified as Independent in the DO (Table 3A). With the expanded model options, the Independent model was selected for 87% of triplets in both the CC and DO. The Complex model was selected for most of the remaining triplets in both studies, but the overall high rate of correct classification indicates little error in the middle variable of the Independent model (genotype data) overall for both studies.

Table 3 Percent of concordant and discordant triplets with each classification under the Three-choice and Expanded model options.

	Three-choice		Expanded	
	CC Liver	DO Liver	CC Liver	DO Liver
(A) Discordant Triplets (%)				
Causal	0.72	1.78	0.41	1.38
Independent	99.17	98.22	87.07	87.35
Reactive	0.10	0.00	0.10	0.00
Complex			12.41	11.17
Other			0.00	0.10
(B) Concordant Triplets (%)				
Causal	10.34	32.13	5.17	15.57
Independent	47.67	38.51	31.54	19.03
Reactive	41.99	29.36	27.61	14.98
Complex			34.64	45.77
Other			1.03	4.65

Concordant triplets. Applying three-choice model selection to concordant triplets, we found that only 10% of triplets in the CC study and 32% of triplets in the DO study were classified as Causal (Table 3B). The remaining triplets were split between the Independent and Reactive models with a larger proportion of Independent classifications, suggesting greater error in the transcript than genotype data. Notably, a greater portion of the CC triplets were Reactive compared to the DO triplets. With the expanded model options, the Causal model was selected for only 5% of CC triplets and 16% of DO triplets, and the Complex model was selected for 35% of CC triplets and 46% of DO triplets. In this setting, partial mediation is possible because the QTL is local to both the transcript and protein. However, we expect that many of the triplets that were classified as Complex are a result of un-modeled measurement error in the transcript data.

Estimated latent correlations. To gain insight into potentially differing error properties between these mouse studies, we estimated the latent correlations of concordant triplets (Figure 7C, see Appendix). Assuming that the Causal model is true, r_{MU} is an estimate of error in the transcript. The frequency of correct classification as a function of r_{MU} was consistent with our simulated data, as can be seen by comparing Figure 5A with Figure 7D. As r_{MU} increased, so did the proportion of triplets classified as Causal with three-choice model selection. The Causal model was not selected until $r_{MU} > 0.6$, and when $r_{MU} > 0.9$ almost 100% of triplets were classified as Causal. Thus, low measurement error in M supports correct inference of the Causal model with both univariate and multi-state X.

The misclassification of concordant triplets as Independent aligns with prior indicators of little genotyping error in both studies (Figure 6B), but why are there more Reactive classifications in the CC than DO? To answer this question, we looked at 254 concordant triplets that were classified as Reactive or Causal in both studies. Assuming the Causal model is true, greater error in the transcript data relative to protein data and a strong causal

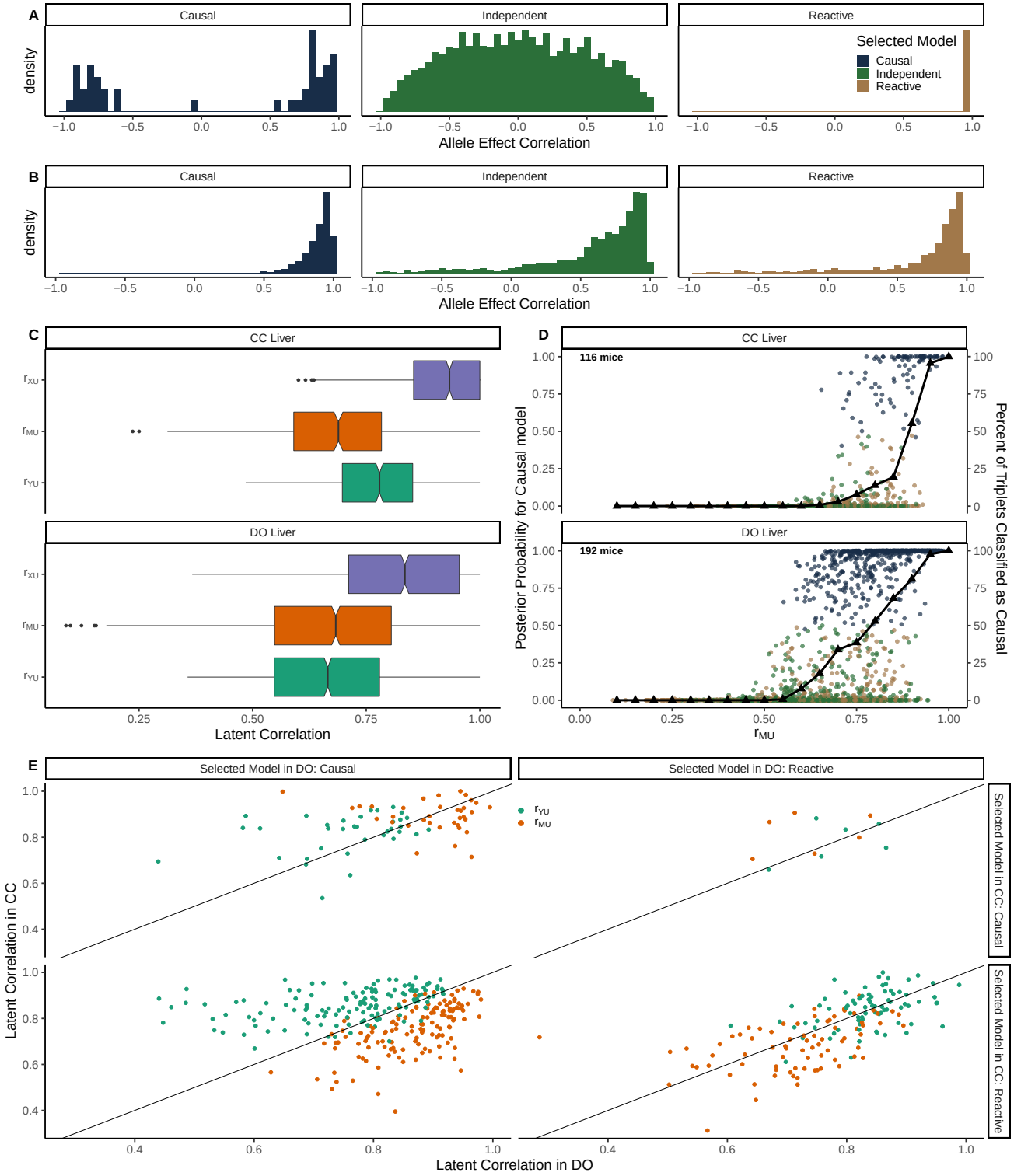

Figure 7 Mediation analysis results for concordant and discordant triplets in DO and CC mouse liver tissue. Distribution of the correlation of eQTL and pQTL allele effects for discordant (A) and concordant (B) triplets, stratified by selected model. (C) Distribution of the estimated latent correlation in the CC (top) and DO (bottom) concordant triplets. (D) Results of three-choice Bayesian model selection for concordant triplets in the CC (top) and DO (bottom). Circle points represent posterior probability for the Causal model as a function of the estimated latent correlation for M, colored by selected model. Black triangle points denote the percent of triplets for which the Causal model was selected at values of r_{MU} rounded to the nearest 0.05. (E) Comparison of estimated latent correlations for M and Y of concordant triplets measured in both the DO and CC and classified as Causal or Reactive, stratified by selected model in each population.

correlation between the transcript and protein data would result in $\rho_{YU} > \rho_{MU}$ and a Reactive classification (Figure 6C). Recall that for the Causal model, ρ_{YU} is equivalent to the product of ρ_{Y^*Y} and $\rho_{Y^*M^*}$ (Table 1). If we further assume that the causal correlations between RNA and protein are similar between the CC and DO (Keele *et al.* 2021), differences in estimated r_{YU} are largely due to measurement error in the proteins (Y). We saw that triplets classified as Causal in the DO and Reactive in the CC generally had a weaker r_{MU} and stronger r_{YU} in the CC than DO (Figure 7E, bottom left plot), consistent with more measurement error in the DO proteins relative to the CC data. For the smaller number of triplets classified as Causal in the CC and Reactive in the DO, the reverse was true (Figure 7E, top right plot). When the selected model was the same in the CC and DO, the error correlations for Y and M were similar across the studies (Figure 7E, diagonal plots). For triplets classified as Causal, values of r_{MU} were generally larger than values of r_{YU}, and vice-versa for triplets classified as Reactive.

By comparing latent correlations for concordant triplets between the CC and DO study, we established support for our hypothesis that there are more Reactive classifications in the CC than the DO because there is less error in the CC proteomics data due to the improved mass-spectrometer technology and study design (Keele *et al.* 2021). Ironically, greater precision in protein measurement results in a higher rate of misclassification of concordant triplets in the CC data. This example highlights the importance of quantifying error in all three variables. A precisely measured Y may be a better surrogate for M^* than is M when there is error in the middle variable (Rockman 2008), leading to inaccurate Reactive classifications.

Case Studies: Diagnosing mediation analysis

Despite the challenges of inferring the correct structure of a three-variable causal system in the presence of measurement error, we can diagnose when mediation analysis is likely to be reliable or not. We selected three examples from DO mice and human cell lines to illustrate scenarios that occur in practice with genetic mediation analysis. In each case, we bootstrapped the observed data using sampling with replacement and estimated data correlations for bootstrap samples. For an assumed causal structure, we used our simulation data to determine ranges of the measurement error model parameters that yield data correlations overlapping the bootstrapped values (see Methods). We then compared the ranges of causal and error correlations for different causal structures and evaluated whether they seem plausible based on the biological context and measurement technologies.

Mediation of distal eQTL in DO kidney. We obtained transcriptome data from kidney tissue of 188 DO mice (Takemon *et al.* 2021) and examined a locus on chromosome 13 where a distal eQTL for *Sfi1* and a local eQTL for *Rsl1* co-map (Figure 8). *Rsl1* is a transcription factor and a biologically plausible negative regulator of *Sfi1* (Krebs *et al.* 2012). The allele effects of the two transcripts were strongly anti-correlated (r = -0.96, p = 7.4e-05). *Sfi1* also had a local eQTL on chromosome 11, which we included as a covariate in the Bayesian model selection procedure.

Bayesian model selection analysis of *Rsl1* as a mediator of the *Sfi1* pQTL with expanded model options assigned the greatest posterior probability to the Complex model (0.846), followed by the Independent model (0.150). Using the three-choice model options, most of the posterior probability was placed on the Independent model (0.977) with a small amount on the Causal model.

Bootstrapping the observed data, we saw that even though there is overlap between the distribution for ρ_{XY} and ρ_{YM} as the weakest data correlation, more configurations with data correlations similar to those observed come from the Independent model (336) compared to the Causal model (41) (Figure 8F-I). If we assume that *Rsl1* mediates *Sfi1*, *i.e.*, the Causal model (Figure 8K), there must be a strong causal correlation between the eQTL genotype and *Rsl1*, little error in the genotype (X), and a moderate amount of error in *Rsl1* (M). Notably, there must be more error in *Rsl1* than the genotype data. We note that *Rsl1* was expressed at very low levels and may therefore have high measurement error (Rockman 2008). In addition, the strong LOD score for *Rsl1* (LOD > 40) is consistent with a strong causal correlation ($\rho_{X^*M^*}$).

Alternatively, if the relationship is Independent (Figure 8J), the ranges of error model parameters are similar to those seen for the Causal model, but with the possibility of a weaker causal correlation between genotype and *Rsl1* and requiring less error in *Rsl1*. The causal and error correlations contributing to ρ_{YU} may span a wide range of values that is similar regardless of the assumed causal structure. We conclude that, allowing for measurement error, the relationship between *Rsl1* and *Sfi1* could be either Independent or Causal.

Epigenetic mediation of gene expression in human lymphoblastoid cell lines. In the next example, we look at epigenetic regulation of transcription in human cell line data. A genetic study identified variants that affected gene expression (eSNPs) and chromatin accessibility (cSNPs) in 63 Lymphoblastoid Cell Lines (LCLs). This included a SNP in an interferon-stimulated response element (ISRE) in the first intron of *SLFN5* that is both an eSNP for *SLFN5* expression and a cSNP for a chromatin peak positioned directly above the ISRE (Degner *et al.* 2012) (Figure 9). The position of the chromatin peak suggests that it mediates *SLFN5*, which is an interferon-regulated gene, by controlling the accessibility of the ISRE to transcription factors (Mavrommatis *et al.* 2013).

Bayesian model selection with expanded model options placed most of the posterior probability on the Complex model (0.82), followed by the Reactive model (0.156). In this setting, partial mediation with direct and indirect effects of the SNP on gene expression is possible because the SNP is local to both the target and candidate mediator. However, it is equally likely that the chromatin state mediates gene expression (Causal model), and the Complex model was selected due to measurement error in the chromatin accessibility data (M). Mediation analysis with three-choice model options selected the Reactive model (posterior probability 0.87), implying that the gene's transcript mediates the local chromatin accessibility, a less plausible scenario.

Bootstrapping the observed data, we saw that more configurations with data correlations similar to those observed come from the Reactive model (799) compared to the Causal model (216) (Figure 9E-H). If we assume that the chromatin peak mediates *SLFN5 i.e.*, the Causal model (Figure 9J), there must be a strong causal correlation between the SNP and chromatin peak ($\rho_{X^*M^*}$) and also between the chromatin peak and the target *SLFN5* ($\rho_{Y^*M^*}$). This is consistent with the strong genetic associations for both the chromatin peak and *SLFN5* ($-\log_{10}(p\text{-value}) > 10$). In addition, there must be little error in the SNP and in *SLFN5* expression, and more measurement error in the chromatin peak than the SNP and *SLFN5*.

Alternatively, if the causal structure is Reactive (Figure 9I),

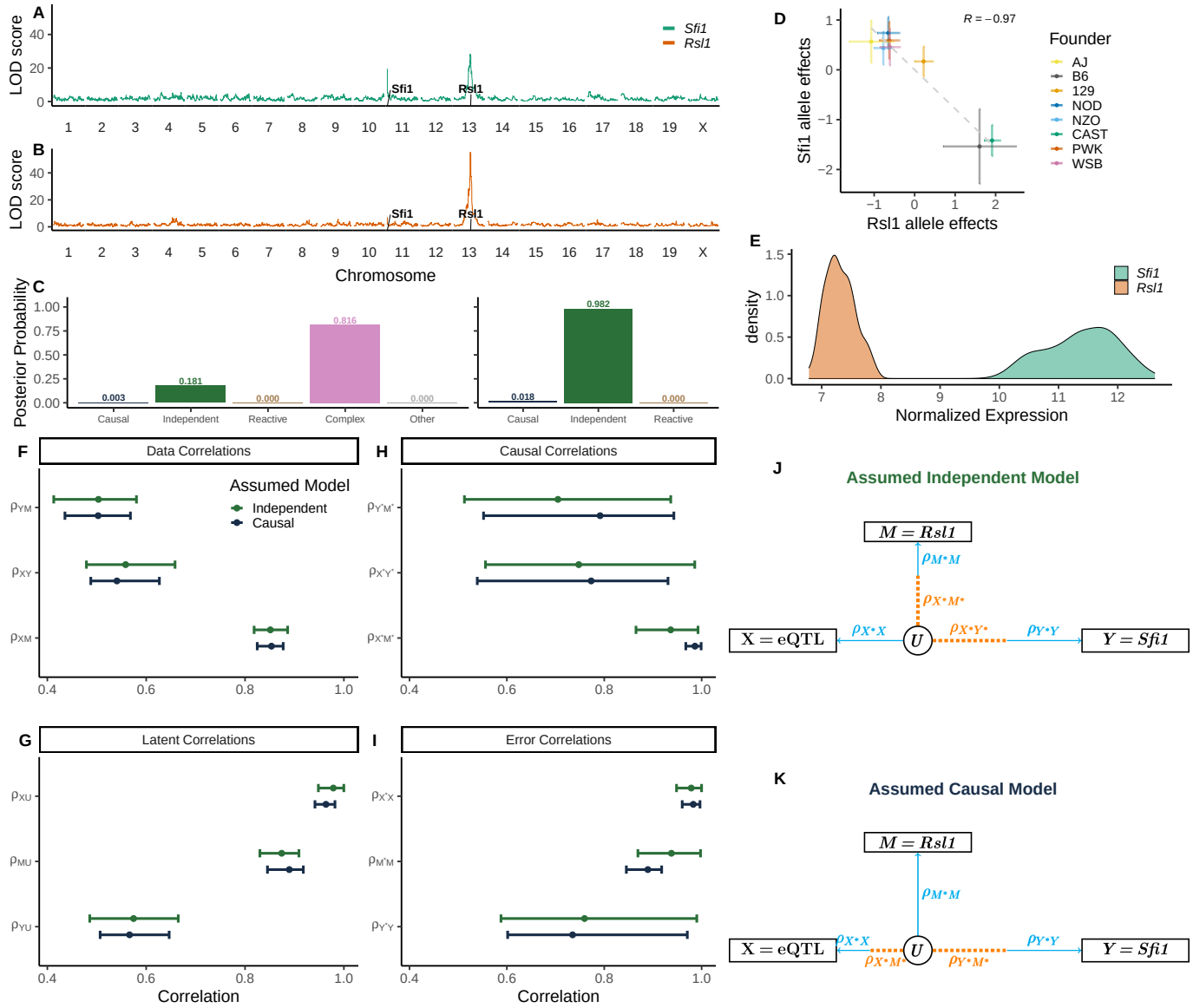

Figure 8 Mediation of *Sfi1* expression in DO kidney tissue. (A) *Sfi1* has a local eQTL on chromosome 11 and a distal eQTL on chromosome 13. (B) The distal eQTL co-localizes with a local eQTL for the transcription factor *Rsl1*. (C) Posterior model probabilities for the structure of the relationship between *Rsl1* and *Sfi1* calculated by Bayesian model selection with the expanded (left) and three-choice (right) model options. (D) Chromosome 13 QTL allele effects for *Rsl1* and *Sfi1*. (E) Variance stabilized transformed expression of *Rsl1* and *Sfi1*. (F-I) Median and 95% highest density interval for distributions of data, latent, causal, and error correlations for measurement error models that could generate the observed data when the causal structure is assumed to be Independent (green) or Causal (blue). DAGs of the latent variable model show relative strengths of causal and error correlations that could produce the data if the assumed model is Independent (J) or Causal (K).

the causal correlation between the SNP and *SLFN5* ($\rho_{X^*Y^*}$) must be strong and there must be little error in the SNP and *SLFN5*. Notably, the strength of the causal correlations $\rho_{X^*M^*}$ and $\rho_{Y^*M^*}$ can span a wider range and are generally weaker compared to the Causal model. Furthermore, the Reactive model requires less error in the chromatin peak and slightly more error in *SLFN5* compared to the Causal model.

Without precise knowledge of the relative measurement errors we cannot come to a firm conclusion. In light of our expectation that chromatin data are noisy and that open chromatin regulates transcript abundance, we suspect that the mediation analysis inference of a Reactive relationship is incorrect and that the true causal structure is either Causal or Complex.

Mediation of distal pQTL in DO Liver. Here we return to the DO liver data (Chick *et al.* 2016) and look at an example of protein to protein mediation (Figure 10). We observed a distal pQTL for TUBG1 that co-maps on chromosome 8 with a local pQTL for NAXD. The chromosome 8 pQTL for TUBG1 also co-maps with a local pQTL for another protein, TUBGCP3. The allele effects for TUBG1 were strongly negatively correlated with those of NAXD (r = -0.89, p = 0.003) and positively correlated with those of TUBGCP3 (r = 0.93, p = 0.0008). TUBGCP3 and TUBG1,

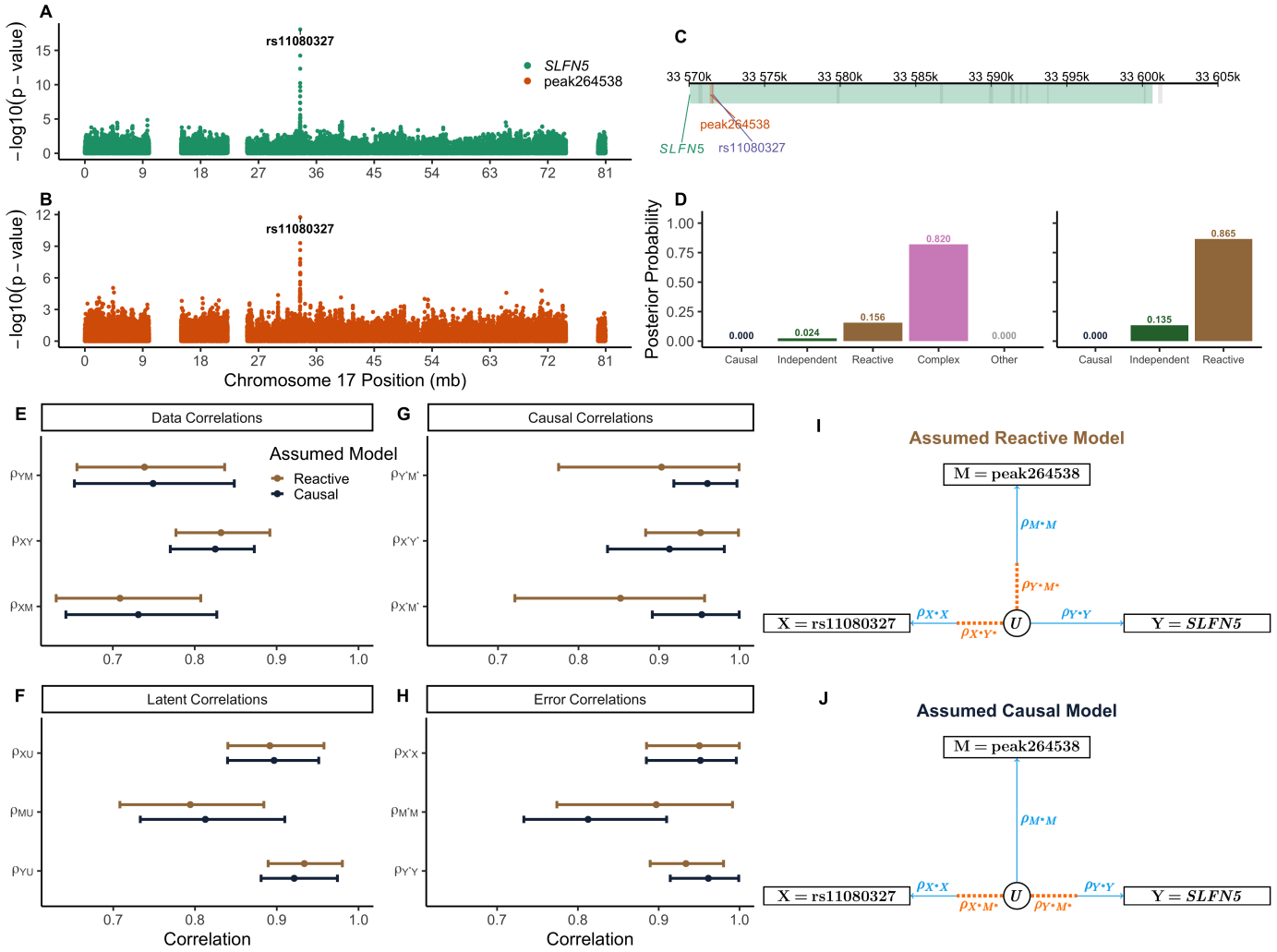

Figure 9 Mediation of *Slfn5* expression by a nearby chromatin peak in LCL data. rs11080327 serves as an eSNP for *SLFN5* (A), and a cSNP for a chromatin peak above the intron (B). (C) The locations of *SLFN5*, rs11080327, and chromatin peak264538 on chromosome 17. (D) Posterior model probabilities for the relationship between the chromatin peak and *SLFN5* calculated by Bayesian model selection with the expanded (left) and three-choice (right) model options. (E-H) Median and 95% highest density interval for distributions of data, latent, causal, and error correlations for measurement error models that could generate the observed data when the causal structure is assumed to be Reactive (gold) or Causal (blue). DAGs of the latent variable model show relative strengths of causal and error correlations that could produce the data if the assumed model is Reactive (I) or Causal (J).

together with a third protein TUBGCP2, form the γ-tubulin small complex (Oakley *et al.* 2015; Farache *et al.* 2018). This functional relationship suggests that TUBGCP3 likely mediates the distal pQTL for TUBG1. In addition, analysis of CC mice (Keele *et al.* 2021) supported TUBGCP3 as a Causal mediator and NAXD as an Independent co-local pQTL.

Bayesian model selection analysis of NAXD as a candidate mediator of the TUBG1 QTL in DO mice placed the greatest posterior probability on the Causal model under both the expanded (0.694) and three-choice (0.978) model options. We also applied Bayesian model selection to test TUBGCP3 as a mediator of the TUBG1 pQTL. Under the expanded model options, the Complex model was selected (posterior probability 0.857), and under the three-choice model options, the Causal model had the greatest posterior probability (0.926).

Bootstrapping the observed data for NAXD, we saw overlap between the distribution for ρ_{XY} and ρ_{YM} as the weakest data correlation, indicating the data may be consistent with the Causal or Independent model. However, more configurations with data correlations similar to those observed come from the Independent model (3,738) compared to the Causal model (1,638) (Figure 10H-K)). Examination of the error model parameter ranges indicated that, if NAXD and TUBG1 are Independent (Figure 10L), there must be a strong causal correlation between the QTL and NAXD ($\rho_{X^*M^*}$), little error in the genotype at the QTL, and little error in NAXD. Alternatively, if NAXD is a mediator of the TUBG1 QTL (Causal model, Figure 10M), the strength of the causal correlation between the QTL and NAXD must be even stronger, with slightly less genotyping error, but more error in NAXD.

Considering the biological evidence supporting TUBGCP3 as a mediator of TUBG1, why does mediation analysis indicate NAXD as the likely mediator? The NAXD QTL (LOD $\approx$ 40) was much stronger than that of TUBG1 (LOD $\approx$ 13). In addition,

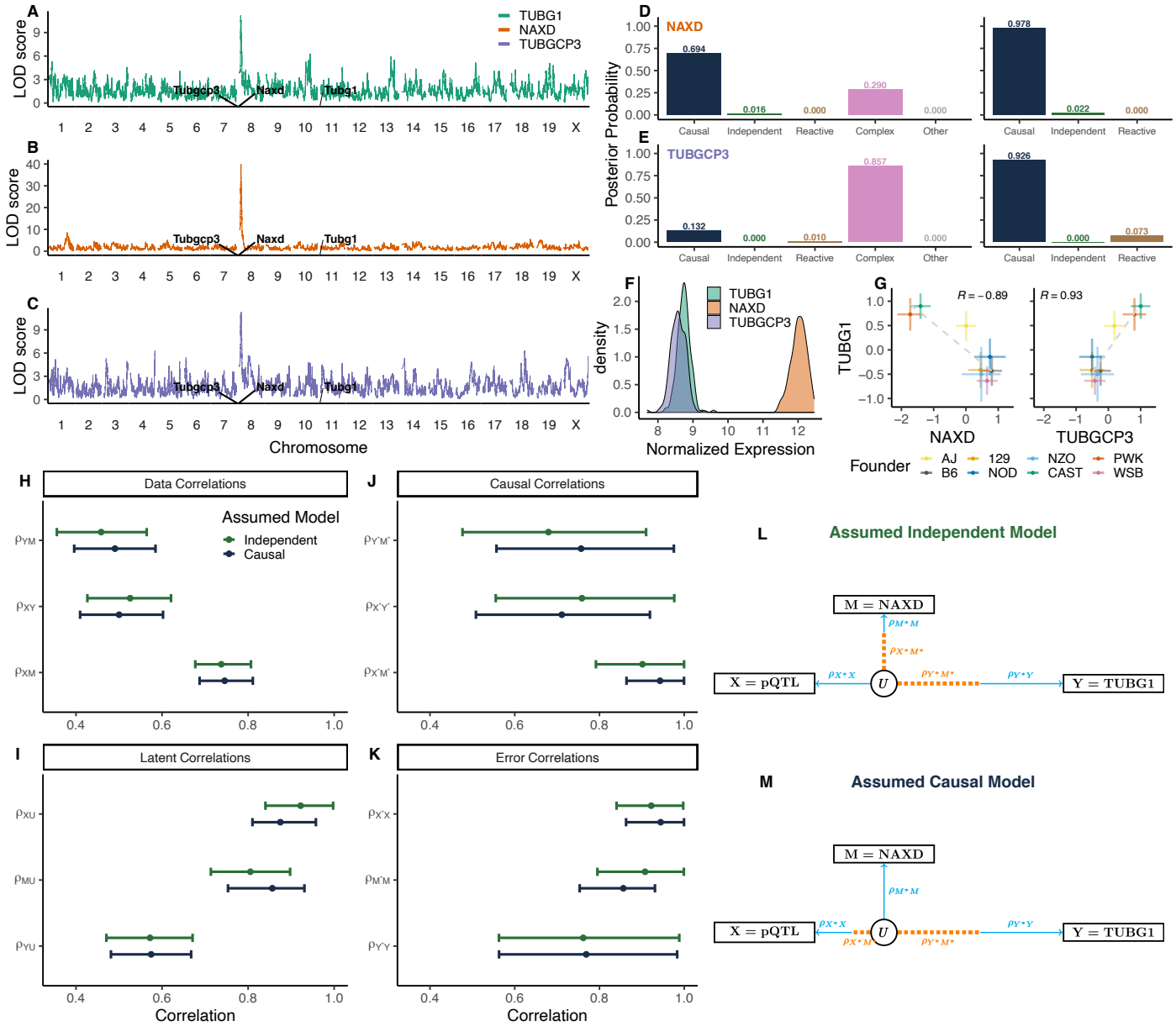

Figure 10 Mediation of TUBG1 protein abundance in DO liver tissue. A distal pQTL for TUBG1 (A) co-localizes with chromosome 8 local pQTLs for NAXD (B) and TUBGCP3 (C). Posterior probabilities for NAXD (D) and TUBGCP3 (E) as candidate mediators of TUBG1, calculated by Bayesian model selection with expanded (left) and three-choice (right) model options. (F) Normalized protein abundance of NAXD, TUBGCP3, and TUBG1. (G) Allele effects for TUBG1 compared to those for NAXD (left) and TUBGCP3 (right). (H-K) Median and 95% highest density interval for distributions of data, latent, causal, and error correlations for measurement error models that could generate the observed data when the causal structure is assumed to be Independent (green) or Causal (blue). DAGs of the latent variable model show relative strengths of causal and error correlations that could produce the data if the assumed model is Independent (L) or Causal (M).

NAXD was more abundant than the target gene TUBG1 (Figure 10F), indicating that TUBG1 may be measured with more error (Rockman 2008). Finally, the alignment of the allele effects of the Chr 8 pQTL on NAXD and TUBGCP3 could be coincidental. The nearly 100% posterior probability for mediation by NAXD creates misleading certainty in the selected model when using three-choice model selection. This is not a flaw in the Bayesian model selection method but is rather due to misspecification of the underlying model which does not account for measurement error.

This example illustrates a common scenario in which mediation analysis indicates that a highly expressed gene with a strong local QTL is identified as a candidate mediator when the true relationship is Independent. In addition, the previous examples illustrate scenarios where it is unclear what the causal structure is without further insight into the relative amounts of measurement error. They demonstrate that mediation inferences may be incorrect due to unbalanced error amongst X, M, and Y; and that data measured with error will be consistent with partial mediation, *i.e.*, the Complex model, even in the absence

of mediation.

Discussion

In this work, we examined the impact of measurement error in mediation analysis. Using theoretical considerations and simulations, we found that in the presence of measurement error, the likelihood of data from any causal relationship of three variables is equivalent to partial mediation (the Complex model). Thus, it is possible to infer partial mediation even in the absence of an indirect effect of X on Y through M. This outcome becomes more likely as sample size increases. This is a property of the model likelihood and thus holds regardless of the inference method used, which is especially concerning for methods such as Sobel's test that focus on detecting the indirect effect.

The measurement error model for three variables is not identifiable, *i.e.*, it is not possible to uniquely estimate each of the model parameters. We introduce a latent variable model in which the products of non-identifiable parameter pairs are estimable and illustrate how it can be used to evaluate the impact of measurement error on mediation analysis. Using the latent variable model, we identify scenarios for which a three-choice model selection strategy (excluding partial mediation) can lead to consistent or inconsistent identification of the causal structure. For a given causal structure, the latent correlations constrain the range of possible causal and error correlations in the measurement error model. Prior knowledge about the measurement errors was critical to diagnosing the mediation results in our case studies. Understanding the relative amount of error in each variable provided insight into the reliability of the mediation analysis results. In particular, if X, Y, and M are measured with similar error, it is possible to infer the correct model.

In genetic mediation analysis, the candidate mediator is often a transcript or protein abundance trait with a local QTL that co-localizes with a QTL for the target. The two most likely causal structures in this scenario are Independent and Causal. A common error in genetic mediation occurs when a gene with a strong local QTL is in linkage disequilibrium (LD) with a true causal mediator that has a weaker effect on the target. The false mediator becomes a surrogate for the genotype due to its strong association with the target's QTL through LD and thus out competes the true mediator. This is illustrated by our case study of NAXD and TUBGCP1.

Another common scenario in genetic mediation of local QTL is the unexpected detection of a Reactive relationship. Assuming the true causal structure is Causal and the three-choice model selection indicates the Reactive model, the mediator may have higher measurement error compared to the target. In our analysis of concordant triplets (genotype, RNA, protein) in CC and DO mice, we observed a higher rate of Reactive classification in the CC mice for which the protein data have reduced measurement error due to improvements in the mass-spec techniques. In another example, Aygün *et al.* (2022) used technical replicates to estimate measurement error in the mediator and target and excluded triplets with unbalanced measurement error for mediation inference. This reduced the rate of false Reactive classifications.

Incorporating a more informative exogenous variable into the system can avoid mediation misclassification. With the DO and CC mouse data, we made use of a multi-state exogenous variable that encodes haplotypes rather than bi-allelic variants. We have previously shown that in the absence of measurement error, compared to a bi-allelic SNP, using multi-state haplotype regression can reduce false detection of mediation and provide stronger evidence of mediation when it is present (Crouse *et al.* 2022). Our analysis of pQTL suggest that this remains true in the presence of measurement error. There were very few cases of discordant triplets (Independent model) mis-classified as Causal. However, in constructing the discordant triplets, we selected candidate mediators based only their position and high LOD scores without considering their allele effects. In practice, if the candidate mediator is selected by scanning all genes in the QTL region, the selected candidate could by chance have allele effects that align with those of the target. Mismatch of the allele effects can rule out the Causal model, but correlated allele effects do not necessarily support a Causal model.

The ideal solution to these challenges would be to measure variables without error, but this is not realistic. We could modify the inference methods to incorporate new modeling assumptions, but any generalization of the three-variable model would not address the fundamental non-identifiability of the causal correlations. As long as we limit ourselves to three variables that are measured once, there is no possibility to model measurement error. It is possible to design experiments that allow for the estimation of measurement error in one or more the variables. As mentioned above, Aygün *et al.* (2022), made use of technical replicates to estimate measurement error. Another avenue is to expand the number of variables in the causal system. In some cases, bidirectional mediation is possible. For example, the relationship between *Rsl1* and *Sfi1* can be resolved by considering the QTL on chromosome 11 (local to *Sfi1*) that does not affect *Rsl1*, supporting *Rsl1* as the mediator of the distal chromosome 13 QTL for *Sfi1*.

Fully modeling a three variable causal system presents some additional challenges, such as addressing the possibility of hidden confounders. In particular, we assume that there are no hidden confounders of $M \rightarrow Y$ that are caused by X (Saunders and Blume 2018). Mendelian randomization (MR) (Katan 1986; Didelez and Sheehan 2007; Richmond *et al.* 2016) is a form of instrument variable analysis, a class of causal inference techniques related to mediation. MR uses an inference method that is robust to the presence of hidden confounders, but it achieves robustness by imposing a different assumption that there is no direct effect between X and Y. Thus MR cannot distinguish between the Independent and Causal mediation models (Crouse *et al.* 2022). For example, MR would fail to distinguish among candidate mediators with correlated allele effects, likely favoring the one with the strongest local QTL because any information M possesses for Y after accounting for X is discarded. The potential impact of measurement error in MR warrants further study.

Genetic mediation analysis is an effective approach to identify the candidate genes and suggest mechanistic explanations for the downstream effects of genetic variation. While the pitfalls of mediation analysis are real, we hope that this deep dive into measurement error in three-variable mediation will lead to more considered application of the method, acknowledging it weaknesses, while not detracting from its utility. It seems possible that incorporating more variables or developing experimental designs that support the estimation of measurement error could mitigate some of these challenges. However, there may be no substitute for independent experimental validation of mediation analysis results.

Methods

Simulations

We simulated data from the measurement error model with a univariate exogenous variable that was normally distributed. For each of the Causal, Independent, and Reactive models, 100,000 configurations were constructed by randomly sampling two causal and three error correlations from a beta distribution: $\rho \sim \text{Beta}(5, 1.25)$. We chose this as a realistic distribution for causal and error correlations with mean 0.8 and 95% highest density interval between 0.5 and 1, allowing for weak correlations but placing greater density on moderate to strong correlations. We compared the distributions of data correlations to those observed in the DO liver data to confirm biological plausibility. Then X^* was sampled from a standard normal distribution and M^* and Y^* were simulated according to the causal correlations using the method described in (Crouse *et al.* 2022) with the effect size calculated as ρ^2. We used the same method to simulate X, M, and Y from their causal counterparts with the desired error correlation.

Mediation Analysis with Bayesian Model Selection

We used Bayesian model selection implemented in the bmediatR R package (Crouse *et al.* 2022) to obtain the posterior probability for the standard mediation model (no measurement error) using either the three-choice or expanded model options. We ran bmediatR with default priors for the effect sizes, and a uniform prior over models. Data were classified according to the model with the greatest posterior probability. With the three-choice model options, the posterior probability was calculated for only the Causal, Independent, and Reactive models. For the expanded model options, the posterior probability was calculated for all models excluding models with an edge $Y \rightarrow M$ that are likelihood equivalent to models with $M \rightarrow Y$. If the greatest posterior probability was not assigned to one of the Causal, Reactive, Independent, or Complex models, the data were classified as non-mediation.

QTL Mapping Analysis

QTL mapping and allele effect estimation in mouse data were done using the qtl2 R package (Broman *et al.* 2019), which fits a linear mixed effect model that accounts for population structure encoded in a genetic relationship matrix, *i.e.*, kinship matrix (Kang *et al.* 2010). Allele effects were estimated as best linear unbiased predictors (BLUPs) to stabilize potentially extreme, lowly observed alleles. Sex, diet, and litter were used as covariates in the DO liver data; sex was used as a covariate in the DO kidney data; and sex was used as a covariate in the CC liver data. QTL in the LCL data were identified by calculating the $-\log_{10}(p - value)$ from regressing chromatin data or gene expression onto the genotype of each SNP compared to a null model with no genotype term.

Using bootstrap sampling to compare real data to simulated data

To identify measurement error model parameters that are consistent with the observed data, we used the boot R package (Davison and Hinkley 1997; Canty and Ripley 2021) to sample with replacement (10,000 times) to generate empirical distributions for the observed data correlation matrix. If X was multivariate we used canonical correlation to estimate ρ_{XM} and ρ_{XY}. Then, we filtered the previously described measurement error model configurations used in data simulations to those that produced data correlations jointly within the range of the bootstrapped distribution.

Data availability

All analyses were performed using the version 4.2.0 of the R statistical programming language (R Core Team 2022). A description of data simulations can be found on page 16, and all data and R code used to generate the results are available at figshare (https://doi.org/10.6084/m9.figshare.20126543).

Data are also available for download and interactive analysis with the QTLViewer webtool (Vincent *et al.* 2022) (https://github.com/churchill-lab/qtlapi) for the DO Liver (https://churchilllab.jax.org/qtlviewer/svenson/DOHFD) and DO Kidney (https://churchilllab.jax.org/qtlviewer/JAC/DOKidney) studies. The individual CC liver data are available in QTLViewer format from figshare (https://doi.org/10.6084/m9.figshare.12818717) at data/qtlviewers/cc_individuals_proteomics_qtlviewer.Rdata. Both genotype (Li *et al.* 2016) and RNA-seq data (Pickrell *et al.* 2010; van de Geijn *et al.* 2015) for Yoruba LCLs are publicly available for download (http://eqtl.uchicago.edu/jointLCL/). The DNase-seq data for the 69 cell lines with RNA-seq data were used as previously processed (Grubert *et al.* 2015).

Funding

This work was support by NIH grants R01 GM076468 (GAC), and F32 GM134599 (GRK).

Conflict of interest

None to declare.

Literature cited

Aygün N, Liang D, Crouse WL, Keele GR, Love MI, Stein JL. 2022. Inferring cell-type-specific causal gene regulatory networks during human neurogenesis. bioRxiv. .

Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of personality and social psychology. 51 6:1173–1182.

Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Śaunak Sen, Yandell BS, Churchill GA. 2019. R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics. 211:495–502.

Byrd R, Lu P, Nocedal J, Zhu C. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal of Scientific Computing. 16:1190–1208.

Canty A, Ripley BD. 2021. *boot: Bootstrap R (S-Plus) Functions*.

Chen LS, Emmert-Streib F, Storey JD. 2007. Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biology. 8:R219.

Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. 2016. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 534:500–505.

Churchill GA, Gatti DM, Munger SC, Svenson KL. 2012. The diversity outbred mouse population. Mammalian Genome. 23:713–718.

Collaborative Cross Consortium. 2012. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics. 190:389–401.

Crouse WL, Keele GR, Gastonguay MS, Churchill GA, Valdar W. 2022. A Bayesian model selection approach to mediation analysis. PLOS Genetics. 18:e1010184.

Davison AC, Hinkley DV. 1997. *Bootstrap Methods and Their Applications*. Cambridge University Press. Cambridge.

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE *et al*. 2012. DNaseI sensitivity QTLs are a major determinant of human expression variation. Nature. 482:390–394.

Didelez V, Sheehan N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. Statistical methods in medical research. 16:309–30.

Farache D, Emorine L, Haren L, Merdes A. 2018. Assembly and regulation of γ-tubulin complexes. Open Biology. 8:170266.

Gatti DM, Svenson KL, Shabalin A, Wu LY, Valdar WW, Simecek P, Goodwin N, Cheng R, Pomp D, Palmer A *et al*. 2014. Quantitative trait locus mapping methods for diversity outbred mice. G3: Genes, Genomes, Genetics. 4:1623–1633.

Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, Greenside P, Srivas R, Phanstiel DH, Pekowska A *et al*. 2015. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 162:1051–1065.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 42:348–354.

Katan M. 1986. Apolipoprotein E isoforms, serum cholesterol, and cancer. The Lancet. 327:507–508.

Keele GR, Zhang T, Pham DT, Vincent M, Bell TA, Hock P, Shaw GD, Paulo JA, Munger SC, Pardo-Manuel de Villena F *et al*. 2021. Regulation of protein abundance in genetically diverse mouse populations. Cell Genomics. 1:100003.

Krebs CJ, Schultz DC, Robins DM. 2012. The KRAB zinc finger protein RSL1 regulates sex- and tissue-specific promoter methylation and dynamic hormone-responsive chromatin configuration. Molecular and cellular biology. 32:3732–3742.

Ledgerwood A, Shrout PE. 2011. The trade-off between accuracy and precision in latent variable models of mediation processes. Journal of personality and social psychology. 101:1174–1188.

Li B, Ritchie MD. 2021. From gwas to gene: Transcriptome-wide association studies and other methods to functionally understand gwas discoveries. Frontiers in Genetics. 12.

Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. 2016. RNA splicing is a primary link between genetic variation and disease. Science. 352:600–604.

Mavrommatis E, Fish EN, Platanias LC. 2013. The schlafen family of proteins and their regulation by interferons. Journal of interferon cytokine research : the official journal of the International Society for Interferon and Cytokine Research. 33:206–210.

Neto EC, Broman AT, Keller MP, Attie AD, Zhang B, Zhu J, Yandell BS. 2013. Modeling Causality for Pairs of Phenotypes in System Genetics. Genetics. 193:1003 LP – 1013.

Oakley BR, Paolillo V, Zheng Y. 2015. γ-Tubulin complexes in microtubule nucleation and beyond. Molecular Biology of the Cell. 26:2957–2962.

Otter T, Pachali MJ, Mayer S, Landwehr JR. 2018. Causal inference using mediation analysis or instrumental variables – full mediation in the absence of conditional independence. Marketing, Zeitschrift fur Forschung und Praxis. 40.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 464:768–772.

Pierce BL, Tong L, Chen LS, Rahaman R, Argos M, Jasmine F, Roy S, Paul-Brutus R, Westra HJ, Franke L *et al*. 2014. Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. PLOS Genetics. 10:e1004818.

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. 2016. Challenges and novel approaches for investigating molecular mediation. Human Molecular Genetics. 25:R149.

Rockman MV. 2008. Reverse engineering the genotype–phenotype map with natural genetic variation. Nature 2008 456:7223. 456:738–744.

Saul MC, Philip VM, Reinholdt LG, Chesler EJ. 2019. High-diversity mouse populations for complex traits.

Saunders CT, Blume JD. 2018. A classical regression framework for mediation analysis: fitting one model to estimate mediation effects. Biostatistics. 19:514–528.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C *et al*. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics. 37:710–717.

Sobel ME. 1982. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociological Methodology. 13:290–312.

Srivastava A, Morgan AP, Najarian ML, Sarsani VK, Sigmon JS, Shorter JR, Kashfeen A, McMullan RC, Williams LH, Giusti-Rodríguez P *et al*. 2017. Genomes of the mouse collaborative cross. Genetics. 206:537–556.

Takemon Y, Chick JM, Gerdes Gyuricza I, Skelly DA, Devuyst O, Gygi SP, Churchill GA, Korstanje R. 2021. Proteomic and transcriptomic profiling reveal different aspects of aging in the kidney. eLife. 10:e62585.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods. 12:1061–1063.

Vincent M, Gerdes Gyuricza I, Keele GR, Gatti DM, Keller MP, Broman KW, Churchill GA. 2022. QTLViewer: an interactive webtool for genetic analysis in the Collaborative Cross and Diversity Outbred mouse populations. G3 Genes | Genomes | Genetics. jkac146.

Appendix

Violation of conditional independence between measured variables.

Conditional independence between measured variables can be violated even when the underlying causal variables are conditionally independent. Suppose the causal structure is a Causal model, *i.e.*, complete mediation, and thus the causal correlations satisfying the conditional independence constraint $\rho_{X^*Y^*} = \rho_{X^*M^*} \cdot \rho_{Y^*M^*}$. In addition, we assume that all of the causal and error correlations are non-zero. Conditional independence between X and Y implies $\rho_{XM} \cdot \rho_{YM} = \rho_{XY}$. This happens when

$$\begin{aligned}
\rho_{XY} - \rho_{XM} \cdot \rho_{YM} &= 0 \\
\rho_{X^*X} \cdot \rho_{X^*Y^*} \cdot \rho_{Y^*Y} - \rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{M^*M} \cdot \rho_{Y^*Y} \cdot \rho_{Y^*M^*} \cdot \rho_{M^*M} &= 0 \\
\rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{Y^*M^*} \cdot \rho_{Y^*Y} \cdot (1 - \rho_{M^*M}{}^2) &= 0 \\
\rho_{M^*M} &= \pm 1.
\end{aligned}$$

Similar algebra shows that if $\rho_{M^*M} = \pm 1$ and the structure of the causal system is Causal then $\rho_{XY} - \rho_{XM} \cdot \rho_{YM} = 0$. Thus, when the structure of the causal system is Causal, conditional independence between measured variables is satisfied if and only if $\rho_{M^*M} = \pm 1$, *i.e.*, there is no measurement error in M. Repeating the above with the constraints for the Independent and Reactive models, we see that measurement error in the middle variable leads to violation of conditional independence between measured variables.

This result is overlooked in most mediation analyses where conditional independence criteria are used to infer complete mediation. For example, Chen *et al.* (2007) introduce the *Causality Equivalence Theorem* in which they prove that the causal relationship $X \to M \to Y$ exists and there are no unmeasured confounders of $M \to Y$ if and only if the following three conditions hold: $X \to Y$, $X \to M$, and $X \perp Y|M$ (conditional independence). The proof of this theorem assumes that all direct causes of each variable are measured without error. However, it is reasonable to assume that measurement error is present in any real setting and conditional independence between the exogenous variable and target will not hold in the measured data even if the structure of the causal relationship is $X \to M \to Y$ (complete mediation).

The measurement error model likelihood.

The log likelihood of the data given a model correlation matrix, Σ, is described by

$$L = -\log|\Sigma| + \mathrm{tr}(\Sigma^{-1}S) \tag{2}$$

where S_{3x3} is the correlation matrix of the observed data. For the Casual measurement error model,

$$\Sigma = \begin{bmatrix} 1 & \rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{M^*M} & \rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{Y^*M^*} \cdot \rho_{Y^*Y} \\ \rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{M^*M} & 1 & \rho_{M^*M} \cdot \rho_{Y^*M^*} \cdot \rho_{Y^*Y} \\ \rho_{X^*X} \cdot \rho_{X^*M^*} \cdot \rho_{Y^*M^*} \cdot \rho_{Y^*Y} & \rho_{M^*M} \cdot \rho_{Y^*M^*} \cdot \rho_{Y^*Y} & 1 \end{bmatrix}.$$

Each combination of measurement error model parameters generates one Σ, but the same Σ may be achieved by different parameter combinations. For example, every entry with ρ_{X^*X} includes the product $\rho_{X^*X} \cdot \rho_{X^*M^*}$. Thus, the values of ρ_{X^*X} and $\rho_{X^*M^*}$ may be swapped, and the resultant Σ will not be changed.

This property holds when the causal structure is Independent or Reactive, indicating that the measurement error model is unidentifiable and the causal and error correlations cannot be uniquely estimated from the data.

However, the latent variable model may also be used to describe the data and thus we can write Σ for any causal structure as follows,

$$\Sigma = \begin{bmatrix} 1 & \rho_{XU} \cdot \rho_{MU} & \rho_{XU} \cdot \rho_{YU} \\ \rho_{XU} \cdot \rho_{MU} & 1 & \rho_{MU} \cdot \rho_{YU} \\ \rho_{XU} \cdot \rho_{YU} & \rho_{MU} \cdot \rho_{YU} & 1 \end{bmatrix}.$$

This formulation of Σ is identifiable and can be used to estimate the latent correlations via maximum likelihood estimation.

We numerically optimized the likelihood in Equation 2 for the observed correlations estimated from standardized data using the "L-BFGS-B" method (Byrd *et al.* 1995) with bounds of (-1, 1) and initial condition 0.5. In the case when X is a multi-state variable, we estimate r_{XY} and r_{XM} with canonical correlations. Doing so results in underestimation of ρ_{XU}, providing an upper bound on the amount of error in X (Figure S7).

Supplemental Materials

Figure S1 Rates of correct and incorrect classifications for data simulated with increasing sample sizes. Sample size is plotted on the x-axis and the percent of data sets with each classification (colored line) is plotted on the y-axis. Columns show results for data simulated with each causal structure. Classification rates are shown for Bayesian model selection with three-choice (top row) and expanded (bottom row) model options. The correct model is denoted with triangle points and the incorrect models are denoted with circles.

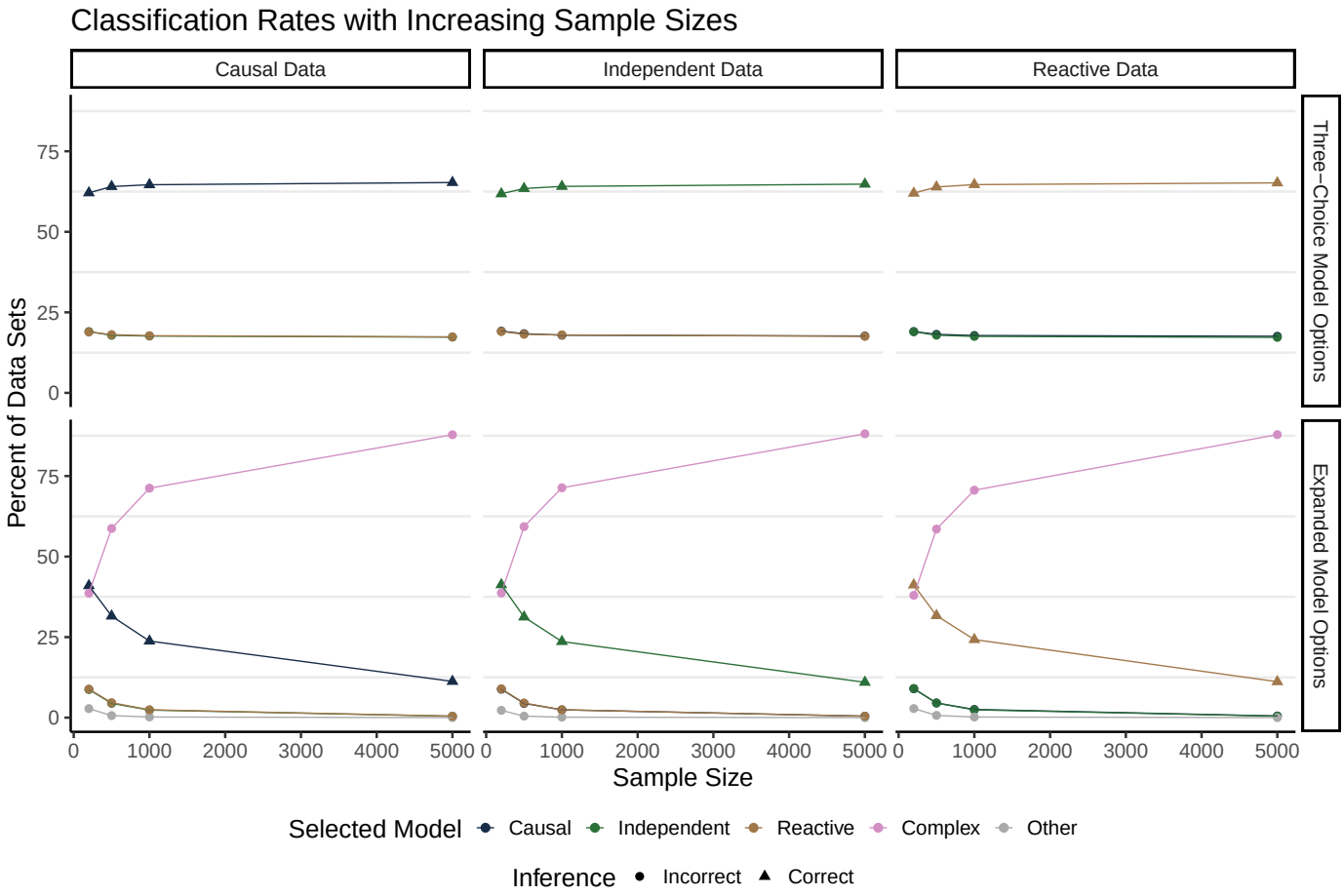

Figure S2 Estimated data correlations from simulations of each causal structure with $N = 5000$ classified with expanded model options. Each row of panels corresponds to simulations of a causal structure with $N = 5000$. Columns correspond to binned values of r_{YM}. The x- and y-axes show r_{XM} and r_{XY}, respectively. Points representing the estimated data correlation from simulated data are colored to indicate the model with the greatest posterior probability from expanded model selection. Shaded regions indicate the three-choice model selection inference rule, and the unshaded region delineates where the correlation matrices are not positive semi-definite. See Figure 4 for results from simulations with $N = 200$.

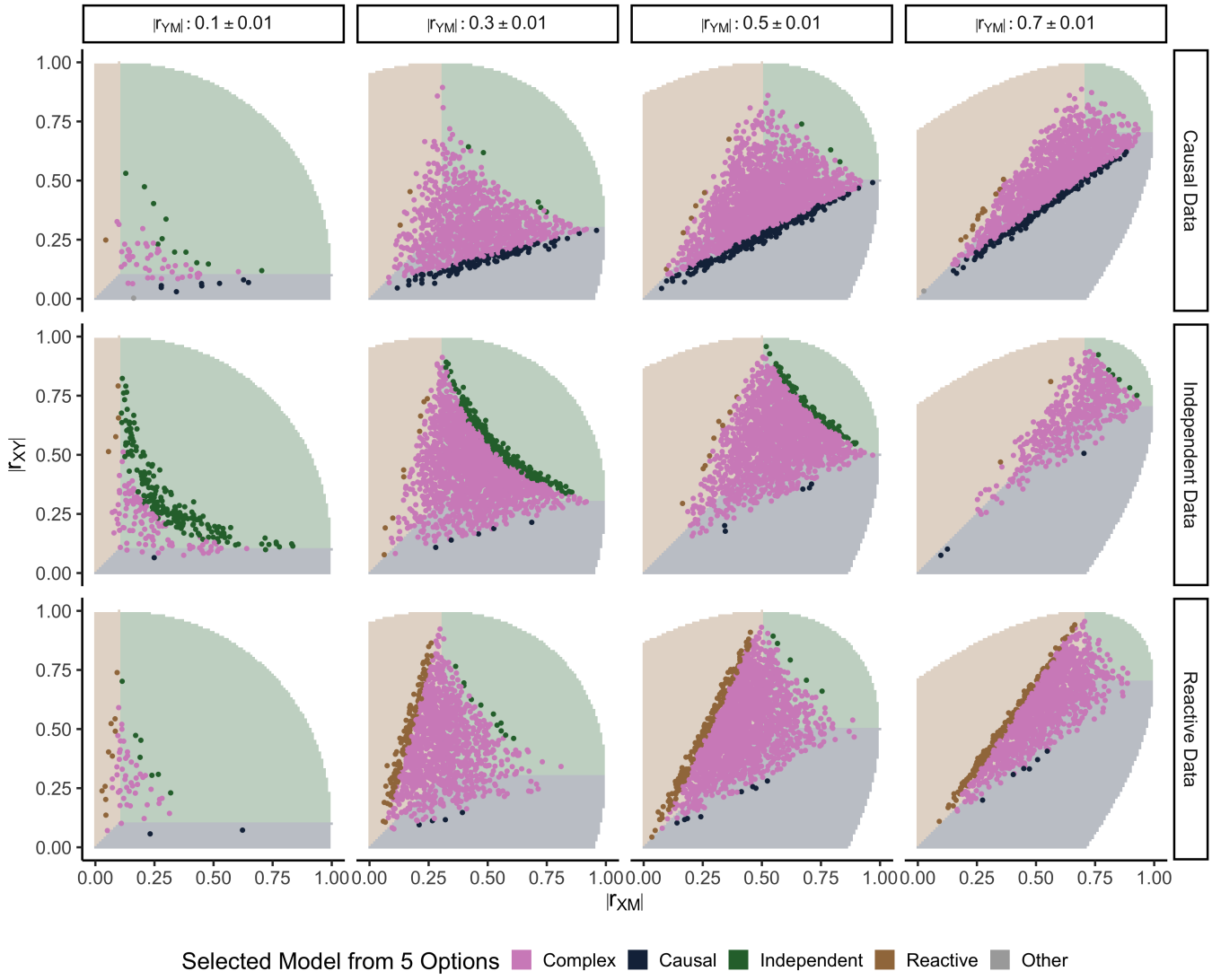

Figure S3 Rates of correct and incorrect classifications for Bayesian model selection with three-choce model options applied to data simulated from the Causal model with varying combinations of latent correlations. ρ_{XU} and ρ_{YU} rounded to the nearest 0.05 are plotted on the x-axis and y-axis, respectively. Increasing values of ρ_{MU}, the latent correlation of the middle variable for the Causal model, rounded to the nearest 0.05 are shown in each column. Shading represents the proportion of data sets simulated at each combination of latent correlations classified as the model labeled at the end of each row. Results are shown for data simulated with with $N = 5000$ (A) and $N = 200$ (B). See Figure S3 for results of model selection with expanded model options.

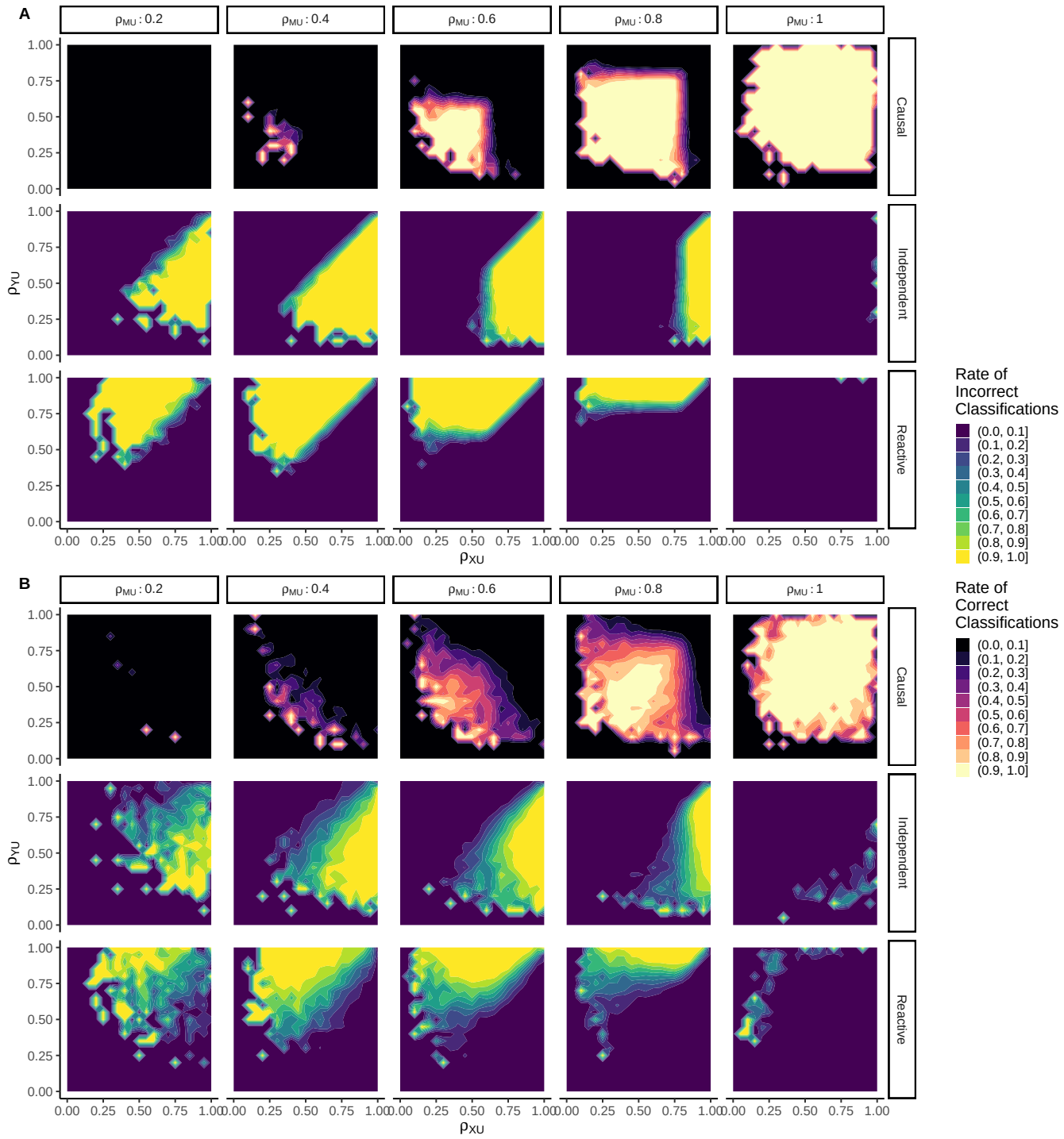

Figure S4 Rates of correct and incorrect classifications for Bayesian model selection with expanded model options applied to data simulated from the Causal model with varying combinations of latent correlations. ρ_{XU} and ρ_{YU} rounded to the nearest 0.05 are plotted on the x-axis and y-axis, respectively. Increasing values of ρ_{MU}, the latent correlation of the middle variable for the Causal model, rounded to the nearest 0.05 are shown in each column. Shading represents the proportion of data sets simulated at each combination of latent correlations classified as the model labeled at the end of each row. Results are shown for data simulated with with $N = 5000$ (A) and $N = 200$ (B). See Figure S3 for results of three-choice model selection.

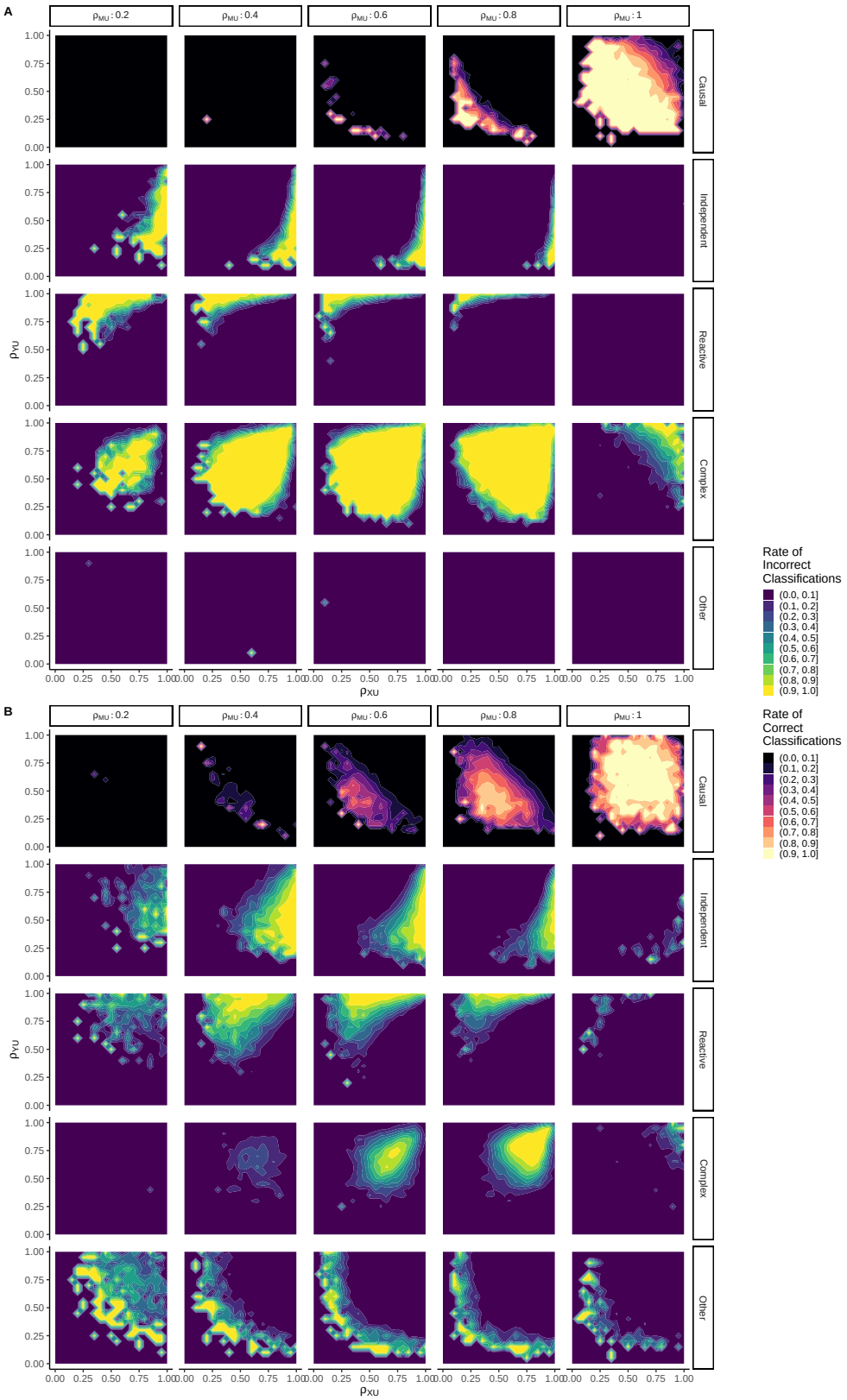

Figure S5 Configurations of the latent variable model representing the Independent model that result in consistent and inconsistent inferences. Configurations of the latent variable model representing the Independent model that result in (A) consistent and (B-C) inconsistent inferences. Figure designed the same way as Fig 6. (A) The correct model is inferred if the latent correlation for X is the strongest (the latent variable arm for X is the shortest). This can be achieved if there is an equal amount of error in all three variables (top right) or if there is less error in X than M and Y (middle left). If X is noisier than M and Y, the correct model may still be inferred if the causal correlations are weak (middle right). The bottom row shows scenarios where M and Y satisfy different configurations. (B) The Causal model is inferred if the latent correlation for M is the strongest. When the causal structure is the Independent model, this will only occur if the error correlation for X is weaker than both the error correlation and causal correlation contributing to the latent correlation for M. The relative magnitude of the causal and error components of the latent correlation for M do not matter as long as their product results in the strongest latent correlation. Similarly, the relative magnitude of the causal and error components of the latent correlation for Y do not matter as long as the latent correlation for Y is not larger than that of M. (C) Shows the analogous scenario to (B) for inferring the Reactive model by swapping M and Y.

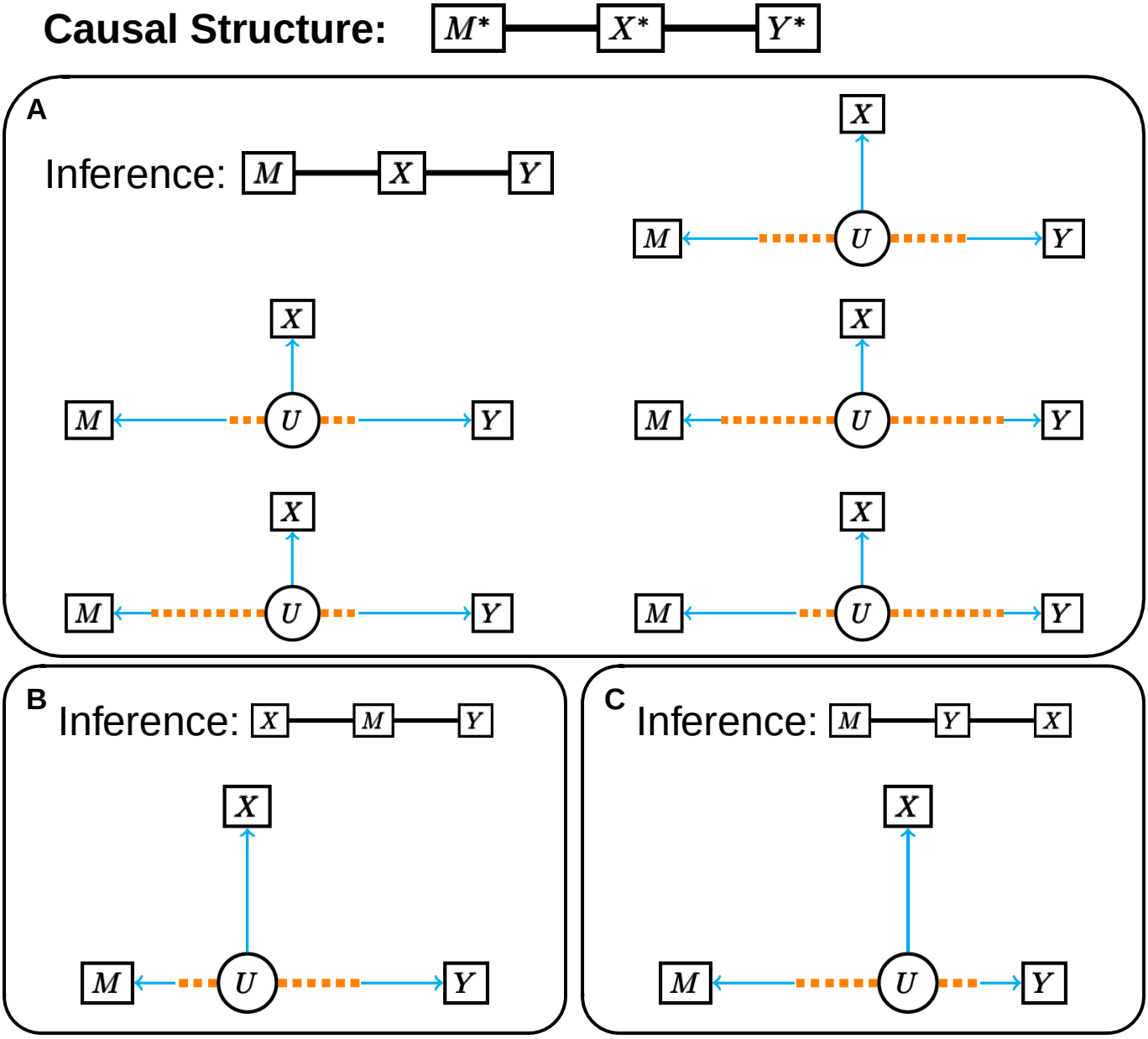

Figure S6 Configurations of the latent variable model representing the Reactive model that result in consistent and inconsistent inferences. Configurations of the latent variable model representing the Reactive model that result in (A) consistent and (B-C) inconsistent inferences. Figure designed the same way as Fig 6. (A) The correct model is inferred if the latent correlation for Y is the strongest (the latent variable arm for Y is the shortest). This can be achieved if there is an equal amount of error in all three variables (top right) or if there is less error in Y than X and M (middle left). If Y is noisier than X and M, the correct model may still be inferred if the causal correlations are weak (middle right). The bottom row shows scenarios where X and M satisfy different configurations. (B) The Independent model is inferred if the latent correlation for X is the strongest. When the causal structure is the Reactive model, this will only occur if the error correlation for Y is weaker than both the error correlation and causal correlation contributing to the latent correlation for X. The relative magnitude of the causal and error components of the latent correlation for X do not matter as long as their product results in the strongest latent correlation. Similarly, the relative magnitude of the causal and error components of the latent correlation for M do not matter as long as the latent correlation for M is not larger than that of X. (C) Shows the analogous scenario to (B) for inferring the Causal model by swapping X and M.

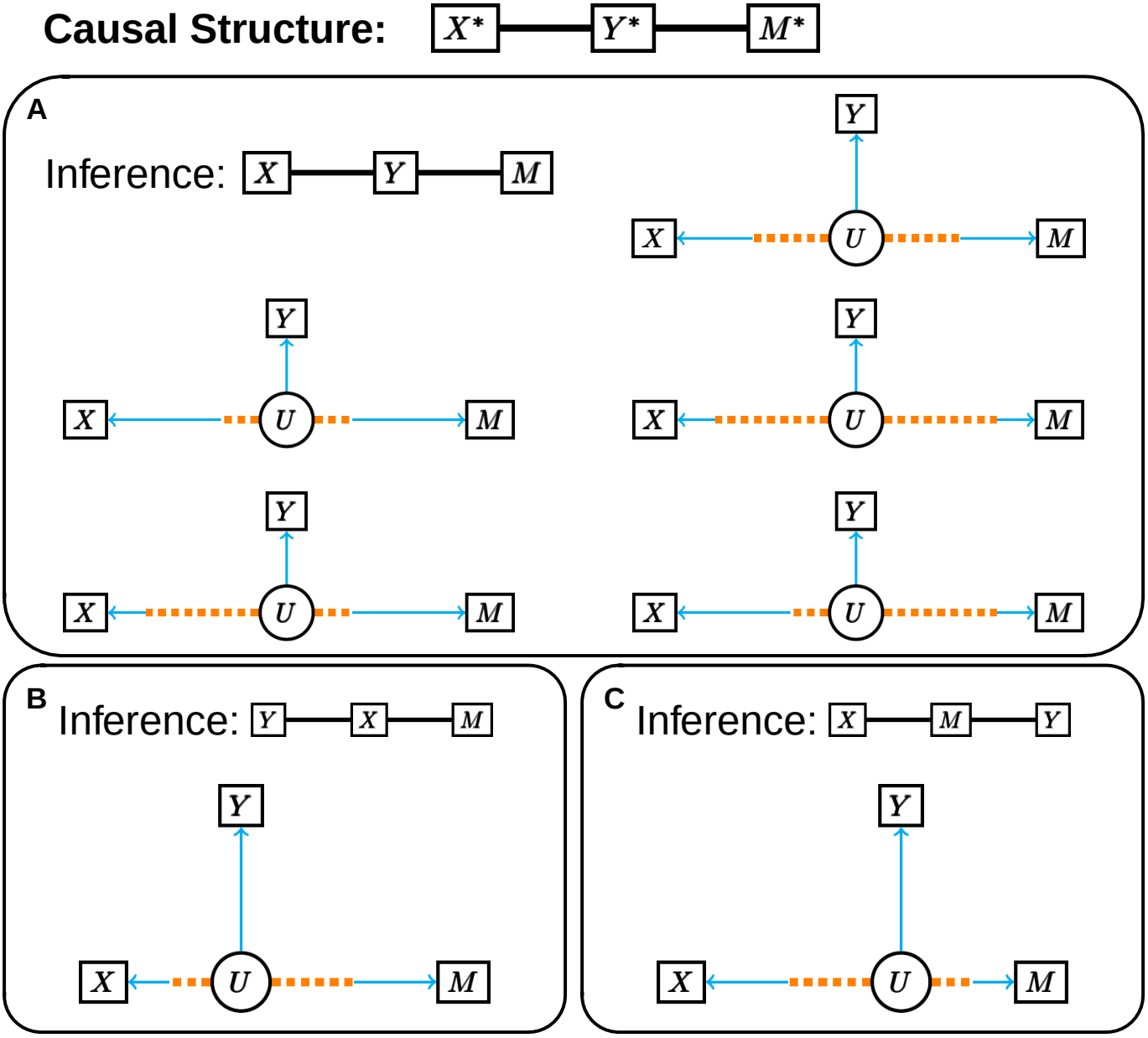

Figure S7 Latent correlations estimated from observed data. (A) Estimated latent correlations for a subset of 1000 simulations of the Causal measurement error model with univariate X. (B) Estimated latent correlations for simulations of the same measurement error models using multi-state X randomly selected from the genotype probabilites from liver tissue of 835 DO mice.

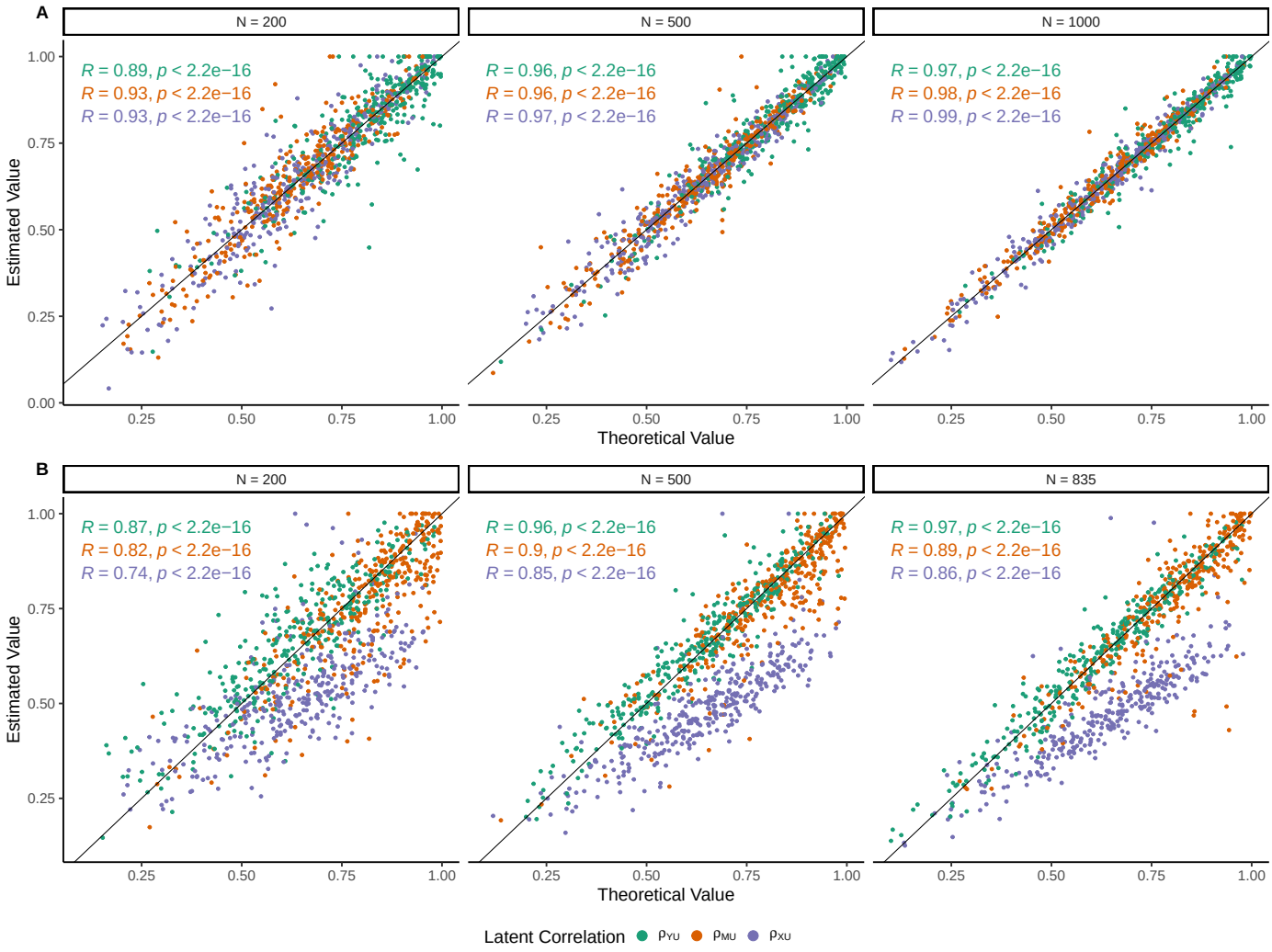

Table S1 Measurement Error Model Configurations. Configurations of the measurement error model that will lead to consistent inferences (diagonal boxes) versus inconsistent inferences (off-diagonal boxes) when classified with three-choice model options.

Causal Structure	**Selected Model**		
	Causal	**Independent**	**Reactive**
Causal	Either $\rho_{M^*M} \geq \rho_{X^*X}$ or $\rho_{M^*M} < \rho_{X^*X}$ & $\rho_{X^*M^*} << 1$ st. $\rho_{M^*M} > \rho_{X^*X} \cdot \rho_{X^*M^*}$ and Either $\rho_{M^*M} \geq \rho_{Y^*Y}$ or $\rho_{M^*M} < \rho_{Y^*Y}$ & $\rho_{Y^*M^*} << 1$ st. $\rho_{M^*M} > \rho_{Y^*Y} \cdot \rho_{Y^*M^*}$	$\rho_{M^*M} << \rho_{X^*X}$ and $\rho_{M^*M} << \rho_{X^*M^*}$ st. $\rho_{M^*M} < \rho_{X^*X} \cdot \rho_{X^*M^*}$ and $\rho_{X^*X} \cdot \rho_{X^*M^*} > \rho_{Y^*Y} \cdot \rho_{Y^*M^*}$	$\rho_{M^*M} << \rho_{Y^*Y}$ and $\rho_{M^*M} << \rho_{Y^*M^*}$ st. $\rho_{M^*M} < \rho_{Y^*Y} \cdot \rho_{Y^*M^*}$ and $\rho_{Y^*Y} \cdot \rho_{Y^*M^*} > \rho_{X^*X} \cdot \rho_{X^*M^*}$
Independent	$\rho_{X^*X} << \rho_{M^*M}$ and $\rho_{X^*X} << \rho_{X^*M^*}$ st. $\rho_{X^*X} < \rho_{M^*M} \cdot \rho_{X^*M^*}$ and $\rho_{M^*M} \cdot \rho_{X^*M^*} > \rho_{Y^*Y} \cdot \rho_{X^*Y^*}$	Either $\rho_{X^*X} \geq \rho_{M^*M}$ or $\rho_{X^*X} < \rho_{M^*M}$ & $\rho_{X^*M^*} << 1$ st. $\rho_{X^*X} > \rho_{M^*M} \cdot \rho_{X^*M^*}$ and Either $\rho_{X^*X} \geq \rho_{Y^*Y}$ or $\rho_{X^*X} < \rho_{Y^*Y}$ & $\rho_{X^*Y^*} << 1$ st. $\rho_{X^*X} > \rho_{Y^*Y} \cdot \rho_{X^*Y^*}$	$\rho_{X^*X} << \rho_{Y^*Y}$ and $\rho_{X^*X} << \rho_{X^*Y^*}$ st. $\rho_{X^*X} < \rho_{Y^*Y} \cdot \rho_{X^*Y^*}$ and $\rho_{Y^*Y} \cdot \rho_{X^*Y^*} > \rho_{M^*M} \cdot \rho_{X^*M^*}$
Reactive	$\rho_{Y^*Y} << \rho_{M^*M}$ and $\rho_{Y^*Y} << \rho_{Y^*M^*}$ st. $\rho_{Y^*Y} < \rho_{M^*M} \cdot \rho_{Y^*M^*}$ and $\rho_{M^*M} \cdot \rho_{Y^*M^*} > \rho_{X^*X} \cdot \rho_{X^*Y^*}$	$\rho_{Y^*Y} << \rho_{X^*X}$ and $\rho_{Y^*Y} << \rho_{X^*Y^*}$ st. $\rho_{Y^*Y} < \rho_{X^*X} \cdot \rho_{X^*Y^*}$ and $\rho_{X^*X} \cdot \rho_{X^*Y^*} > \rho_{M^*M} \cdot \rho_{Y^*M^*}$	Either $\rho_{Y^*Y} \geq \rho_{M^*M}$ or $\rho_{Y^*Y} < \rho_{M^*M}$ & $\rho_{Y^*M^*} << 1$ st. $\rho_{Y^*Y} > \rho_{M^*M} \cdot \rho_{Y^*M^*}$ and Either $\rho_{Y^*Y} \geq \rho_{X^*X}$ or $\rho_{Y^*Y} < \rho_{X^*X}$ & $\rho_{X^*Y^*} << 1$ st. $\rho_{Y^*Y} > \rho_{X^*X} \cdot \rho_{X^*Y^*}$