END-TO-END MEDIATION ANALYSIS AT SCALE: FROM HIGH-DIMENSIONAL CAUSAL DISCOVERY TO PATH-LEVEL EFFECTS, WITH A HUMAN PHENOTYPE APPLICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal mediation analysis reveals how an exposure affects an outcome, but requires the mediator and the variables to adjust for to be specified in advance; with thousands of variables, both must be inferred from a causal structure learned from the data. Learned structures are uncertain and may disagree on what to adjust for, and existing pipelines face three obstacles at this scale: structure learning grows quadratically with the number of variables, averaging estimates across structures targets no single adjustment, and estimators must commit to an unknown functional form. We propose **Ch**annel-**A**dditive **R**obust **M**ediation (**CHARM**), an end-to-end framework that estimates mediation effects from incomplete, high-dimensional data. CHARM searches for structure only among each variable's strongest partners and registered role pairs, so its cost grows linearly; fixes, before estimation, one adjustment set valid under every structure supporting a path, and withholds the path when none exists; and estimates every path with one flexible model that learns how the mediator acts on the outcome, rather than fixing its form in advance. In simulations, screened discovery matches the accuracy of the full search at a fraction of its cost on data with 10% of values missing; under a complete public tier order, CHARM recovers the target path and estimates its effects accurately up to ten thousand variables, never issues an invalid adjustment set where the structures disagree, and stays accurate across linear and nonlinear mechanisms. On the Human Phenotype Project, it recovers mediation chains that survive multiple-testing correction and reproduce reported biology, such as *Prevotella* acting on blood glucose through branched-chain amino acids.

1 INTRODUCTION

Scientific questions often concern not only whether a factor affects an outcome, but also by what mechanism. Causal mediation analysis addresses them by decomposing the total effect of this factor, the exposure, into an indirect effect through an intermediate variable, the mediator, and a direct effect through all other pathways (Robins & Greenland, 1992; Pearl, 2001). This requires specifying in advance the mediator and the variables to adjust for, the adjustment set (Imai et al., 2010). In high-dimensional data with thousands of measured variables, such manual specification is infeasible. The mediator and the adjustment set must instead be inferred from a causal structure learned from data, a map of which variables influence which (Kalisch & Bühlmann, 2007; Zheng et al., 2018). Recent end-to-end methods estimate effects on such learned structures (Geffner et al., 2022; Emezue et al., 2023), and some target mediation directly (Cai et al., 2021; Li et al., 2025). Learned structures, however, are uncertain: finite, incomplete observational data can be compatible with multiple structures that may disagree on which variables to adjust for. The central challenge addressed here is to obtain reliable mediation effects at scale without relying on the correctness of any single learned structure.

Existing approaches face three obstacles at this scale. 1) **Limited scalability**: most structure learners examine every pair of variables or optimise a dense matrix over all pairs, so their cost grows at least quadratically with the number of variables (Zheng et al., 2018; Lopez et al., 2022), and filling in missing values must scale too. 2) **Unresolved structural uncertainty**: when the data fit several structures, existing pipelines estimate the effect under each structure separately and average the estimates (Emezue et al., 2023; Li et al., 2025). Because each structure may prescribe a different adjustment set, the averaged estimates are computed under different adjustments, and the average is not the effect under any single one of them. Even if each structure has its own valid set, no set need be valid under all of them, so the set most likely to be valid (Hu & van der Pas, 2025) may

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

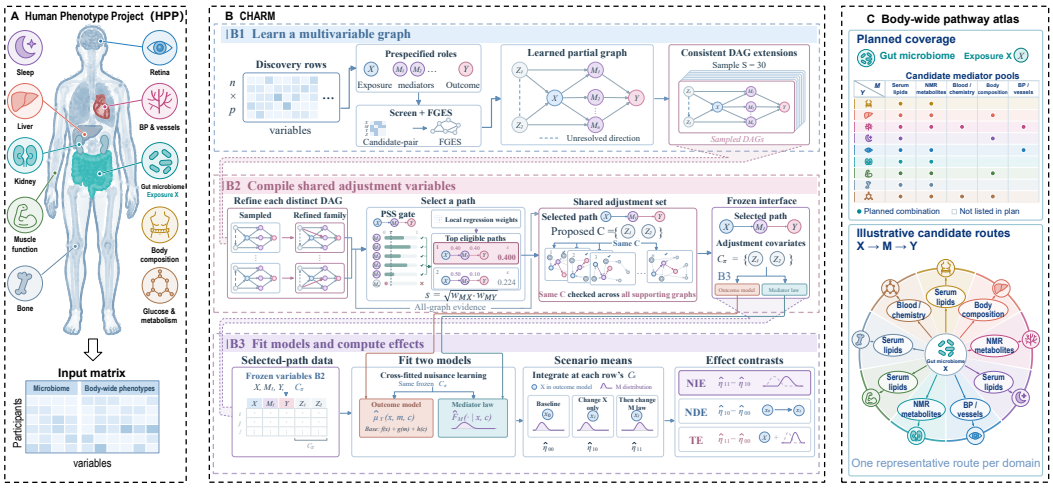


Figure 1: Overview of CHARM on the HPP study. **(A)** Input matrix, with the gut microbiome as the exposure X. **(B)** The three stages, one per obstacle of Section 1: **(B1)** screened discovery, **(B2)** common adjustment-set compilation and **(B3)** channel-additive estimation. **(C)** Scope of the HPP application: the five candidate mediator pools (M) and nine outcome domains (Y) of the registered plan, with one representative route $X \to M \to Y$ per domain.

still fail under some, and a well-fitting structure can still yield a wrong effect (Dang et al., 2026). 3) **Unknown functional form**: estimators must commit to a form for relationships of unknown shape. Penalized high-dimensional mediation methods let the exposure and mediators act linearly on the outcome, adjusting for prespecified covariates either linearly (Zhang et al., 2016) or flexibly (Jones et al., 2025). Such restrictive forms can fail severely when these relationships are nonlinear, whereas fully flexible models lose accuracy when they are simple (Section 4.2).

We introduce **Ch**annel-**A**dditive **R**obust **M**ediation (**CHARM**), an end-to-end framework that turns an incomplete observational record (Figure 1A) into path-level effect estimates through three stages that address the three obstacles in turn (Figure 1B); Figure 1C shows the range of mediators and outcomes in our application to the Human Phenotype Project.

1 **Screened discovery** (Figure 1**B1**): each variable is paired only with a fixed number of its most correlated partners and with the pairs required by its registered role, and a standard structure search (Ramsey et al., 2017) runs only on these pairs, so the search domain grows at most linearly with the number of variables; missing values are likewise filled in from a fixed number of predictors per variable. The search output is then expanded into a family of plausible structures.

2 **Common adjustment-set compilation** (Figure 1**B2**): CHARM retains the paths that recur across the family and, for each path, compiles a single adjustment set that is valid under every structure supporting the path. This set is fixed before any effect is estimated, so that the target no longer varies across structures. When no such set exists, the path is withheld rather than estimated, and we characterise when a single structure admits no such set.

3 **Channel-additive estimation** (Figure 1**B3**): a single fixed model, shared by all paths, expresses the outcome as a linear exposure term plus flexible functions of the mediator and the adjustment set, capturing nonlinear mediator effects without losing accuracy on linear ones or selecting a model per path; the direct, indirect and total effects then follow from this model and a fitted mediator model.

Our contributions are as follows. First, we scale mediation analysis to thousands of variables: the structure search examines a set of variable pairs whose size grows only linearly with the number of variables, and, under a complete public tier order, CHARM recovers the target path and estimates its effects accurately up to ten thousand variables (Section 4). Second, CHARM handles disagreement among learned structures before estimation, and in simulations it never issues an invalid adjustment set where they disagree (Section 4.3). For each path, it adjusts for a set of variables valid in every learned structure containing the path, and does not report the path if none exists. Third, one estimator

applies to linear and nonlinear mechanisms and is accurate in every one we tested (Section 4.2). Finally, we apply CHARM to the Human Phenotype Project, where it recovers mediation chains that survive multiple-testing correction and reproduce reported biology (Section 5).

2 PRELIMINARIES AND PROBLEM SETUP

Estimands. Write X for the exposure, Y for the outcome, $\mathcal{M}$ for the set of candidate mediators, and C_π for the fixed path-level adjustment set compiled in Section 3.4. Let x_0 and x_1 be two target exposure levels; the labels 0 and 1 do not imply that X is binary. With $M(x)$ the potential mediator and $Y(x, m)$ the potential outcome, define the four nested counterfactual means

$$\eta_{ab} := \mathbb{E}\big[Y\{x_a, M(x_b)\}\big], \qquad a, b \in \{0, 1\}, \tag{1}$$

and take $\mathrm{NDE} = \eta_{10} - \eta_{00}$, $\mathrm{NIE} = \eta_{11} - \eta_{10}$ and $\mathrm{TE} = \eta_{11} - \eta_{00} = \mathrm{NDE} + \mathrm{NIE}$. These are the standard pure natural direct and total natural indirect effects (Robins & Greenland, 1992; Pearl, 2001). For a selected mediator M, NIE collects every route from X to Y through M and in general coincides with the effect of the two-edge path $X \to M \to Y$ only when M is the only intermediate, so mediator-specific effects need not add. Effects attached to a single named route are path-specific effects (Avin et al., 2005) and lie outside our scope.

Maintained conditions. Identification relative to a fixed adjustment set C rests on five conditions, stated formally as (A1)–(A5) in Appendix A together with the identifying formula: (A1) pre-exposure invariance of the set, (A2) consistency, (A3) single-world exchangeability of the mediator and the outcome given C, (A4) the cross-world condition and (A5) positivity of the exposure and of nested mediator support. The compiler of Section 3.4 targets the graphical part of (A3), and (A4) is untestable and assumed; the missing-data conditions are in Section 3.5.

3 METHOD: CHARM

3.1 CHARM FRAMEWORK OVERVIEW

CHARM keeps the three-stage decomposition of SIGMA (Li et al., 2025) and changes what each stage hands to the next. From an observational, possibly incomplete high-dimensional record, screened score-based discovery yields a graph family on which stable paths $\pi = (X, M, Y)$ are selected (Sections 3.2–3.3). The bridge compiles the family into one fixed adjustment set $\widehat{C}_\pi$ per path, certified across the path-supporting graphs and withheld when the issuance conditions are unmet (Section 3.4). One prespecified channel-additive base class, extended only when a pre-registered interaction test flags the path, then supplies a single cross-fitted plug-in analysis under one missing-data mode, with no per-graph estimate to average (Sections 3.5–3.6). The record specified for each path separates numerical, graphical, causal, data-law and integrity status (Section 3.7). The formal scope is a single mediator with the estimands of Section 2.

3.2 SCREENED SCORE-BASED STRUCTURE DISCOVERY

Discovery searches one table $\widetilde{D}$, completed once before screening with the completion fitted on discovery rows alone. End-to-end runs split rows first; incomplete inputs use K-MICE (Appendix G.1).

The screen pools each variable's 32 strongest correlation partners and the registered role pairs into an undirected search domain $\mathcal{E}_{\mathrm{allow}}$. Membership permits an edge without forcing its presence or direction, and with a fixed partner budget and mediator list the domain grows at most linearly in p.

The FGES implementation in Tetrad (Ramsey et al., 2017; 2018) searches this domain using degenerate-Gaussian BIC (Andrews et al., 2019) for mixed-type data, under registered role constraints $X \prec m \prec Y$ for $m \in \mathcal{M}$. Degree and penalty settings are fixed per run (Table 5). The output $\widehat{H}$ is a partially directed graph retaining unresolved adjacencies.

End-to-end runs additionally use a complete public five-tier order and admit baseline-to-role pairs, while Table 1 uses neither. Auxiliary local-regression weights only rank paths and alternative adjustment proposals. Both configurations and the scoring conventions are in Appendices G.2 and G.3.

3.3 GRAPH FAMILY AND STABLE PATH SELECTION

From $\widehat{H}$ we independently draw S consistent DAG extensions $G_s^{(0)}$ with replacement (Wienöbst et al., 2021), and each distinct graph undergoes bounded, score-improving parent-set refinement for registered mediators and outcomes. Writing $\mathcal{R}$ for this refinement, the resulting family and path stability score (PSS) are

$$G_s = \mathcal{R}\big(G_s^{(0)}\big), \qquad \widehat{\mathrm{PSS}}(\pi) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\big\{X \to M \in G_s,\ M \to Y \in G_s\big\}. \tag{2}$$

Draw multiplicities are retained. Refinement can change the skeleton and colliders, so G_s need not extend $\widehat{H}$. PSS measures recurrence under this construction, not a posterior or causal probability.

Registered paths with $\widehat{\mathrm{PSS}}(\pi) \geq \tau = 0.2$ qualify, with τ fixed before effect estimation; end-to-end runs select the highest-ranked qualifying path, and if none qualifies no adjustment set is issued. The supporting graphs, those containing both path edges, determine certification and panel diagnostics, whereas node evidence uses all S graphs. Construction and ranking details are in Appendix G.3.

3.4 IDENTIFICATION-AWARE BRIDGE VIA A COMMON ADJUSTMENT SET

Path recurrence does not determine an adjustment set. Even when each supporting graph admits one, a set valid across them need not exist. CHARM compiles one adjustment set, shared by all nuisance models and fixed before effect estimation (Appendix B).

A set C is operationally eligible on G if it contains only recorded variables outside $\{X, M, Y\}$ and no descendant of X. Graphical compatibility additionally requires the three back-door conditions

$$\mathrm{Cert}_G^{\mathrm{op}}(C;\pi) = 1 \iff \mathrm{Elig}_G^{\mathrm{op}}(C;\pi) = 1,\ \ X \perp_d M \mid C \text{ in } G_{\underline{X}},\ \ X \perp_d Y \mid C \text{ in } G_{\underline{X}},\ \ M \perp_d Y \mid X, C \text{ in } G_{\underline{M}} \tag{3}$$

Here $G_{\underline{Z}}$ removes all edges leaving Z. On a fully observed true DAG containing $X \to M \to Y$, the oracle set B_{GT} of all non-descendants of X outside the path satisfies these conditions exactly when the obstruction set $\mathcal{L}_\pi(G)$ of descendants of X that are ancestors of M and, avoiding M, of Y is empty (Proposition 1). A nonempty obstruction defeats every eligible set.

The compiler pools ancestor, parent and descendant evidence across the S graphs, scores candidate sets by their certification rates on two fixed panels, each holding half of the draws, and reads neither the true graph nor any effect estimate; end-to-end runs use a pool cap $Q = 12$ and a set-size cap $K = 2$. Its pooling, ranking, alternative-proposal and ambiguity rules are in Appendix B.

An adjustment set is issued only if it is nonempty, has at most K members, passes equation 3 on every supporting graph in both panels, and each panel contains at least three supporting graphs; unless an alternative proposal is used, a tie among candidates with identical graph evidence does not withhold the set (Appendix B). A refusal, written $\perp$, is not the explicitly estimated $\varnothing$ comparison set; the path is recorded as not run with its reason, and the planned denominator retains the case. Refusal does not establish that no valid adjustment exists.

Both panels enter selection, so their rates describe in-sample compatibility, not independent validation. Certification is relative to the supplied graphs, and the causal label stays withheld by default (Appendix B). For an issued set, write $C_\pi := \widehat{C}_\pi$; only (X, M, Y, C_π) is passed on to estimation.

3.5 CROSS-FITTED CHANNEL-ADDITIVE NUISANCE LEARNING

For each path with an issued C_π we estimate the mediator law $F_M(m \mid x, c)$ and the outcome mean $\mu_Y(x, m, c)$, both conditioning on the same C_π that identification uses. Outer K_{cf}-fold cross-fitting over the fixed estimation sample $\mathcal{I}_{\mathrm{est}}$ makes every prediction entering an effect out of fold. K-MICE completions are fitted on outer training rows alone and only applied to the held-out fold; neither the discovery imputer nor $\widetilde{D}$ reaches the effect stage.

The outcome model uses the prespecified channel-additive base class

$$\mu_{Y,\mathcal{F}}(x, m, c) = f(x) + g(m) + h(c), \tag{4}$$

with $f(x) = \beta_X x$ and flexible g and h. The mediator learner and the training settings of both models are fixed before any contrast is evaluated. Additivity is a working restriction, not an identification assumption, and the base class excludes exposure–mediator interaction. The only departure is an

interaction extension, which adds a term $x\, g_{XM}(m)$ when a pre-registered test flags the path before outer-fold fitting; a flagged path carries a warning label (Appendix C).

Estimation uses exactly one missing-data mode per path: fully observed, observed-row weighting or fold-wise completion, with the weighting conditions in Appendix C. A path moves to fold-wise completion as a whole, never fold by fold, and only when valid fold-specific K-MICE completions are available; otherwise it is recorded as not run and never upgraded.

For a binary exposure its propensity only diagnoses overlap, and under observed-row weighting the observation propensity only weights nuisance training; neither weights the outer mean, taken over the observed values of C_π or, under fold-wise completion, their completed values. Learners, the bias identity, missingness conditions, switching rules and support diagnostics are in Appendix C.

3.6 Counterfactual integration and natural-effect estimation

Let $\mathcal{J}_{\text{main}} = \{(0,0),(1,0),(1,1)\}$ index the three nested means of the decomposition. Over the estimation sample $\mathcal{I}_{\text{est}}$ the cross-fitted nuisances enter the mediation g-formula row by row,

$$\widehat{\gamma}_{ab,i} = \int \widehat{\mu}_Y^{(-k(i))}(x_a, m, C_{\pi,i})\, d\widehat{F}_M^{(-k(i))}(m \mid x_b, C_{\pi,i}), \qquad \widehat{\eta}_{ab} = \frac{1}{n_{\text{est}}} \sum_{i \in \mathcal{I}_{\text{est}}} \widehat{\gamma}_{ab,i}, \tag{5}$$

with no graph index in either quantity. The plug-in contrasts $\widehat{\text{NDE}}$, $\widehat{\text{NIE}}$ and $\widehat{\text{TE}}$ formed from $\widehat{\eta}_{ab}$ are g-functional contrasts; whether they read as natural effects is the causal status of Section 3.7.

Evaluating the integral. A discrete mediator makes the integral a finite sum, enumerated exactly. A continuous one is integrated in full, since for a non-affine channel the conditional mean cannot in general replace the integral. Its law is represented as $M = \nu_M(X, C_\pi) + \varepsilon$ with a residual law that does not vary with (X, C_π), a working restriction rather than a general conditional-distribution learner, and a path relying on it carries the working-distribution label. Every contribution indexed by $\mathcal{J}_{\text{main}}$ must be finite on the estimation sample. The quadrature bank, with one probability measure serving both exposure arms, and the support and variance diagnostics are in Appendix D.

3.7 Path-level effect records and status reporting

For every selected path the record holds the effect vector, its variance diagnostic, the fixed C_π, $\widehat{\text{PSS}}(\pi)$ and the information needed to reconstruct the analysis, with nine status fields (Appendix E); the runs of Section 4 record the adjustment set, the missing-data mode, the support rules, the integration details, the estimation sample and the effect status. The fields vary independently: a finite g-functional may refer to the completed data and carry a withheld causal label at once.

In the record specification, absence of a number has exactly two states. A path is *not run* when the estimator did not execute, as when the bridge issued no adjustment set, and *undefined* only when the target, inputs, row population or a required integral cannot be uniquely determined or finitely evaluated. Support, family and integrity warnings label an estimate and never alter its value. Appendix E gives the possible states, the recorded reasons, the integrity grades and the denominators.

4 Experiments

We ask four questions. **Q1.** How do structure-recovery accuracy and computational cost change as the dimension grows (Section 4.1)? **Q2.** Given the true path and a valid adjustment set, does the fixed estimator stay accurate across outcome mechanisms (Section 4.2)? **Q3.** When the adjustment set is compiled from learned structure, does estimation stay accurate, and does the set pass the admissibility check on the true graph (Section 4.3)? **Q4.** Does CHARM recover reported biology on real data (Section 5)? Baseline rosters and metric definitions are in Appendix F.

4.1 High-dimensional structure recovery

Setting. Each scale draws synthetic DAGs at $n{=}10^4$ with 10% of values missing at random, and $R{=}10$ replicates up to $p{=}2000$ and $R{=}3$ at $p{=}5000$. All four methods receive the same K-MICE-completed table and role tier and are scored against the true graph over all $\binom{p}{2}$ pairs, so a true edge

Table 1: **High-dimensional structure discovery.** Accuracy is scored over all $\binom{p}{2}$ pairs, so a screened-out true edge counts as a miss. Entries are replicate means, $\pm$ standard error where shown, and bold marks the best value in each column at each scale. *N.SHD* counts extra, missing, reversed and uncommitted edges per true edge. *Abst.* is the share of detected true edges left undirected, *CAcc.* the accuracy of committed directions, and *Pooled Path-F1* the directed F1 over the role subgraph $\{X, M, Y\}$. *Pairs searched* is the share of $\binom{p}{2}$ searched. Wall time and recorded memory are replicate summaries that exclude the shared completion, and the p=5000 scale is planned at three replicates because of its cost. Definitions are in Appendix F.

Problem scale		Discovery quality				Cost		
Scale (p;n;R)	Method	Adj-F1 ↑	N.SHD ↓	Orient. F1 ↑ / Abst. ↓ / CAcc. ↑	Pooled Path-F1 ↑	Pairs searched ↓	Wall time ↓	Mem. (GiB) ↓
Panel A: accuracy and cost at the configurations of Table 5								
500; 10k; 10	CHARM	**.773 ± .006**	**.519 ± .014**	**.738** / 3.7% / .971	**.738 ± .053**	**9.0%**	534 s	**1.5**
0.125M pairs	FGES-Full (Ramsey et al., 2017)	.766 ± .006	.543 ± .014	.730 / 2.7% / .965	.691 ± .065	100%	1008 s	6.5
	PC-Stable (Colombo & Maathuis, 2014)	.579 ± .006	1.658 ± .025	.341 / **1.8%** / .586	.646 ± .035	100%	**79** s	5.4
	BOSS (Andrews et al., 2023)	.579 ± .004	.675 ± .004	.550 / 13.1% / **.990**	.405 ± .082	100%	97 s	**1.4**
1000; 10k; 10	CHARM	**.774 ± .004**	**.521 ± .008**	**.739** / 4.4% / .973	.778 ± .050	**4.5%**	2300 s	1.6
0.500M pairs	FGES-Full	.767 ± .004	.544 ± .008	.732 / 3.0% / .967	**.789 ± .050**	100%	7245 s	6.0
	PC-Stable	.532 ± .004	1.963 ± .021	.306 / **1.4%** / .573	.686 ± .031	100%	**310** s	15.9
	BOSS	.579 ± .004	.678 ± .006	.549 / 13.6% / **.989**	.474 ± .065	100%	515 s	**1.4**
2000; 10k; 10	CHARM	**.768 ± .005**	**.528 ± .010**	**.737** / 4.1% / .977	**.767 ± .034**	**2.2%**	2.77 h	73.3
1.999M pairs	FGES-Full	*exhausted memory, no replicate completed*						
	PC-Stable	.441 ± .006	2.549 ± .025	.248 / **1.5%** / .560	.629 ± .065	100%	**0.58** h	90.3
	BOSS	.535 ± .004	.738 ± .004	.502 / 14.4% / **.979**	.636 ± .082	100%	1.16 h	**8.9**
Panel B: top scale, three replicates								
5000; 10k; 3	CHARM	**.762 ± .008**	**.508 ± .016**	**.733** / 5.0% / .982	**.755 ± .076**	**0.9%**	12.57 h	196.0
12.498M pairs	FGES-Full	*exhausted memory, no replicate completed*						
	PC-Stable	.321 ± .011	4.163 ± .045	.175 / **0.8%** / .540	.563 ± .064	100%	**12.05** h	84.5
	BOSS	.564 ± .007	.700 ± .012	.535 / 13.7% / **.992**	.750 ± .150	100%	18.86 h	**4.0**

the screen drops counts as a miss. Up to p=2000, FGES-Full is the search of CHARM without its screen, which isolates the effect of screening; PC-Stable and BOSS run at their tuned configurations (Table 5). Protocols, metrics and the time and memory conventions are in Appendix F.

Recovery and comparison. As the pair space grows a hundredfold, CHARM's adjacency F1 stays at .762–.774. Where FGES-Full completes, it matches CHARM, so the screen costs no accuracy, while PC-Stable and BOSS fall well below both at every scale. CHARM also has the highest orientation F1 at every scale and a path F1 that is the highest or within one standard error of it; BOSS orients more precisely but leaves more edges undirected, and PC-Stable does the reverse (Table 1).

Computational cost. Up to p=2000, PC-Stable and BOSS run faster than CHARM but fall well short of its accuracy. FGES-Full matches that accuracy only at a much higher cost in time and memory, and exhausts memory from p=2000 on. By confining the search to the screened pairs, CHARM keeps full-search accuracy at a fraction of that cost and is the only method that is both accurate and completes at every scale (Table 1, Appendix F).

4.2 EFFECT ACCURACY ACROSS MECHANISM FAMILIES

Effect-only benchmark. We hold the inputs fixed by supplying every method with the true path, a valid adjustment set and the same data, so that the panel isolates effect estimation rather than discovery. The benchmark uses p=100, n_{total}=10^4, n_{effect}=5000, R=10, complete data, and seven outcome-mechanism families under continuous and binary exposure. The families share one structural core and differ only in the shape of the mediator–outcome channel g, with no exposure–mediator interaction, so the panel tests shape adaptation inside the additive class. Every method returns finite NIE, NDE and TE estimates on every replicate, so misspecification is reflected in estimation error. For each family we report NIE and NDE errors separately, and summarise a method by the *relative mean error* L_{rm} (in percent), the average of the two relative absolute errors,

$$L_{\text{rm}} = \tfrac{1}{2}\left(\frac{|\widehat{\text{NIE}} - \text{NIE}|}{|\text{NIE}|} + \frac{|\widehat{\text{NDE}} - \text{NDE}|}{|\text{NDE}|} \right). \tag{6}$$

Structured flexibility across mechanisms. Commitment risk is asymmetric across the evaluated additive mechanisms. Parametric g-computation and PL-DML approach the oracle-form diagnostic on the linear family but reach worst-family L_{rm} of 44.1–57.3% across exposure profiles. CHARM's maxima are 7.1% under continuous and 9.6% under binary exposure, each within 0.6 percentage points of the corresponding oracle-form maximum, and its equal-family means are the lowest among

Table 2: **Effect-estimation accuracy benchmark across seven mechanism families.** Per-family cells give **NIE MAE / NDE MAE** over R=10 replicates and *Macro* is their equal-family mean. *Mean* and *Worst* are the equal-family mean and the largest family value of L_{rm} from equation 6, with that family in parentheses. The *Oracle-form* row is a finite-sample reference diagnostic under the family-specific correct functional form. Baselines and metric definitions are in Appendix F.

Method	Per-family NIE MAE / NDE MAE							Macro NIE / NDE	Mean L_{rm}	Worst L_{rm}
	Linear	Sin	Tanh	Ramp	RBF	Softplus	Poly2			
Panel A: Continuous exposure										
Oracle-form diagnostic	*.007/.017*	*.008/.012*	*.006/.016*	*.007/.016*	*.007/.015*	*.008/.017*	*.010/.010*	*.008/.015*	*5.63%*	*6.5% (Softplus)*
CHARM	.009/.019	.010/.013	.008/.016	.008/.016	.009/.015	.008/.019	.011/.010	**.009/.015**	**6.12%**	**7.1%** (Linear)
CV-selected (van der Laan et al., 2007)	.008/.017	.010/.016	.013/.021	.012/.021	.010/.019	.014/.023	.015/.012	.012/.019	7.59%	9.4% (Softplus)
Spline (Wood, 2017)	.013/.023	.010/.016	.011/.020	.012/.021	.010/.019	.014/.023	.015/.012	.012/.019	7.93%	9.4% (Softplus)
MLP (Farrell et al., 2021)	.018/.045	.018/.022	.019/.034	.019/.035	.025/.032	.012/.044	.018/.016	.018/.033	12.74%	15.5% (Linear)
GPS weighting (Huber et al., 2020)	.016/.042	.074/.107	.034/.061	.031/.061	.036/.065	.018/.040	.056/.082	.038/.066	25.84%	45.3% (Sin)
Parametric g-comp (Robins, 1986)	.007/.017	.091/.093	.030/.033	.031/.032	.049/.043	.046/.033	.094/.024	.050/.039	22.32%	46.1% (Sin)
PL-DML (Robinson, 1988; Chernozhukov et al., 2018)	.010/.017	.085/.091	.026/.031	.027/.031	.045/.040	.049/.031	.095/.023	.048/.038	21.47%	44.1% (Sin)
Panel B: Binary exposure										
Oracle-form diagnostic	*.009/.028*	*.012/.022*	*.009/.028*	*.010/.027*	*.010/.028*	*.009/.029*	*.014/.019*	*.010/.026*	*8.99%*	*9.4% (RBF)*
CHARM	.009/.028	.011/.022	.009/.027	.011/.028	.008/.027	.008/.029	.014/.019	**.010**/.026	**8.94%**	**9.6%** (Ramp)
CV-selected	.008/.028	.013/.022	.010/.027	.010/.028	.010/.027	.010/.029	.014/.019	.011/.026	9.15%	9.8% (Softplus)
Spline	.009/.029	.013/.022	.010/.027	.010/.028	.010/.027	.010/.029	.014/.019	.011/.026	9.20%	9.8% (Softplus)
MLP	.025/.030	.017/.023	.022/.025	.024/.028	.023/.028	.028/.028	.027/.010	.024/**.025**	12.11%	14.1% (Softplus)
Mediation DML (Farbmacher et al., 2022)	.038/.025	.045/.019	.043/.024	.041/.024	.042/.024	.034/.024	.043/.035	.041/.025	16.50%	19.6% (Poly2)
Parametric g-comp	.008/.028	.114/.115	.039/.046	.041/.048	.063/.067	.046/.030	.094/.062	.058/.057	28.64%	57.3% (Sin)
PL-DML	.011/.029	.108/.109	.035/.046	.037/.047	.058/.063	.048/.031	.095/.062	.056/.055	27.81%	54.3% (Sin)

Table 3: **End-to-end composition after high-dimensional discovery. (a)** Target counts replicates recovering the exact target path; Graph-adm. additionally requires true-graph adjustment validity and no selected exposure descendants, both over the planned R. L_{rm} is the relative mean error of equation 6 and MaxAE the largest absolute error across replicates and estimands. Compiled and supplied adjustment sets coincide in every replicate. The second block reruns the design of the first row on ten further seeds; its MAR row adds 10% missingness at random to all variables but the exposure. **(b)** Stress tests in which the learned graph family may disagree about the adjustment set: sets issued and valid, withheld, or issued and invalid. Protocols and diagnostics are in Appendix I.

(a) End-to-end campaigns

Campaign				Gates		Accuracy		
Data	p	n_{total}	R	Target	Graph-adm.	MAE (ind./dir./total)	L_{rm}	MaxAE
complete	500	10k	40	40/40	40/40	.0080/.0102/.0104	4.54%	.0432
complete	5k	10k	5	5/5	5/5	.0043/.0051/.0094	2.36%	.0164
complete	10k	20k	5	5/5	5/5	.0029/.0081/.0107	2.76%	.0196
Design of the first row on ten further seeds, paired: complete vs. 10% *MAR*								
complete	500	10k	10	10/10	10/10	.0096/.0107/.0170	5.08%	.0333
10% MAR	500	10k	10	10/10	10/10	.0137/.0159/.0296	7.41%	.0587

(b) Stress tests

Design		Adjustment sets		
Graph family	n	Valid	Withheld	Invalid
A common valid set exists				
disagreeing	148	147	1	0
agreeing	32	26	0	6
Boundary design				
no common set	60	0	60	0
control, common set	40	40	0	0

deployable methods. The unrestricted MLP has higher mean error in both profiles and fares worst on the linear family under continuous exposure. CV-selected and Spline come closest but stay slightly above CHARM under both exposures, and mediation DML and GPS weighting trail it clearly (Appendix H.1). In the evaluated settings, one fixed class suffices without per-path model selection.

4.3 CLOSING THE NUMERICAL LOOP AFTER HIGH-DIMENSIONAL DISCOVERY

End-to-end setting. Neither Table 1 nor Table 2 shows that discovery and estimation compose. Every replicate here starts from a p-dimensional matrix, runs discovery, selects the top-ranked learned path, compiles one adjustment set with no user-specified C_π, and hands it to the effect stage of Section 3.5. The main campaign has $(p, n_{\mathrm{total}}, R) = (500, 10^4, 40)$ in four registered strata of ten replicates each, which keep the target path and its effects fixed and add no stress variables (ordinary), near-instruments that reach the targets only through X, descendants of an exposure–confounder collider, or strong exposure descendants. Discovery runs under the end-to-end settings of Section 3.2, which place every variable in a complete public five-tier order, and the bridge reads S=30 graphs under the settings of Section 3.4. The campaign is fresh and complete-data, isolating the composition from imputation error. Its seeds are disjoint from a registered list of earlier seeds, its plan was fixed before any data were generated, and the true graphs and effects were revealed only after every replicate had finished. Two further campaigns repeat this design without stratification, at $(p, n_{\mathrm{total}}, R) = (5000, 10^4, 5)$ and $(10^4, 2 \times 10^4, 5)$.

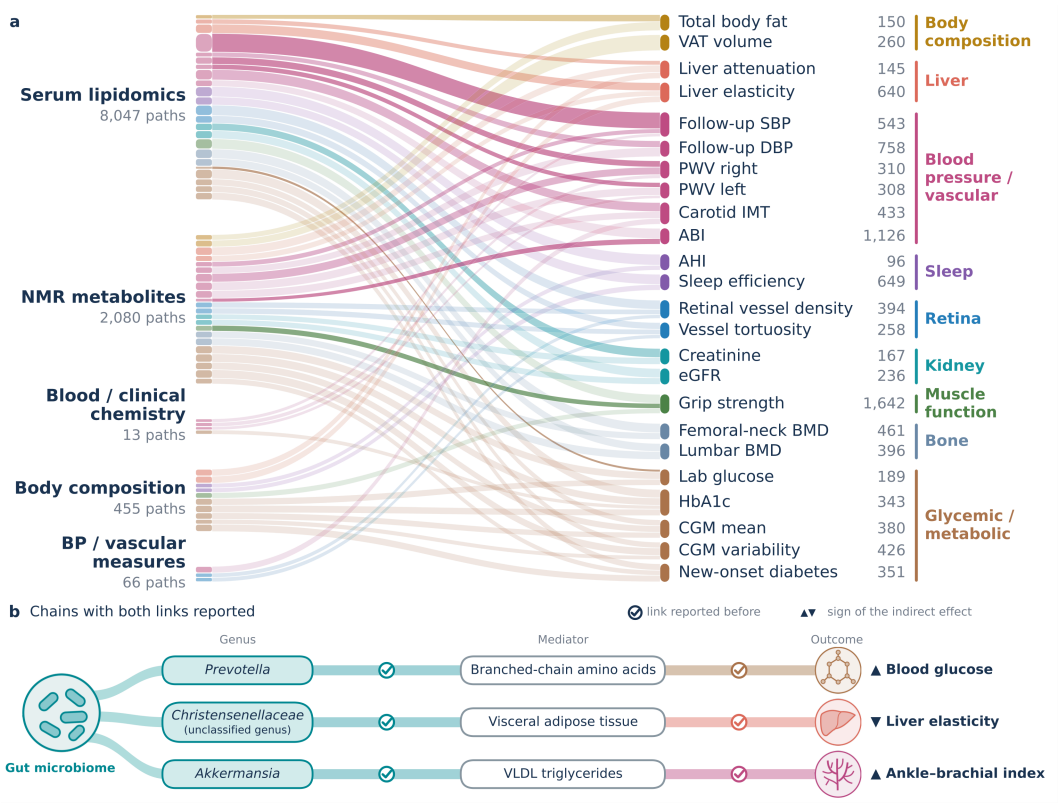


Figure 2: The registered HPP campaign and the chains it recovers. **(a)** The registered routes, from the five mediator pools to the twenty-four outcomes in nine body domains, with the paths estimated. **(b)** Three chains whose links and effect signs agree with published studies; estimates are in Table 19.

Closing the loop. All 40 replicates reached estimation and returned finite contrasts, and each selected the exact target path. The compiled adjustment set passes the joint admissibility check on the revealed true graph in all 40, which puts the lower end of the two-sided 95% Wilson interval at 0.912, above the registered 0.80 threshold, and each of the four registered strata passes 10/10. Both larger campaigns repeat this outcome, 5/5 on each gate.

Accuracy matches the supplied adjustment set. The compiled and supplied arms use the same adjustment set in every replicate and return identical estimates under the matched estimation configuration, so their MAEs coincide exactly for every estimand. The compiled arm reaches an L_{rm} of 4.54%, against 46.01% for the unadjusted comparison (Appendix I).

Missing data and stress tests. Ten further seeds reproduce the main campaign (Table 3a); with 10% MAR missingness both gates still pass, and L_{rm} rises from 5.08% to 7.41%, far below the 44.05% without adjustment. In stress tests (Table 3b), the compiler issues no invalid set where the graph family disagrees and withholds one whenever no common set exists; it corrects 35 of 41 single-graph errors, introduces none, and errs only when every graph misses the same weak edge.

5 Real-data application: the Human Phenotype Project

The Human Phenotype Project (HPP)[1] is a prospective cohort that profiles each participant across body systems (Shilo et al., 2021; Reicher et al., 2025). We take the gut microbiome, profiled by shotgun metagenomics at the baseline visit, as the exposure; serum lipidomics, NMR metabolomics, blood tests, body composition and blood-pressure and vascular measures as candidate mediators; and outcomes in nine body domains, from vascular and retinal imaging to sleep and continuous

[1]Data access and documentation: `https://knowledgebase.pheno.ai/`.

glucose monitoring. Mediators are taken within 45 days of the stool sample and outcomes at a follow-up about two years later, with the baseline outcome as a covariate. Participants are split into a discovery half, used to learn the structure and select paths, and an effect half, used only to estimate them. Because estimation requires the follow-up outcome, a path uses 370–2,560 participants in discovery and 120–1,050 (median 280) in estimation (Appendix J).

Before estimation, we registered 64 routes, each pairing one mediator pool with one outcome (Figure 2a; Appendix J.1); within a route, discovery selects the candidate paths from a genus through one mediator to the outcome. For each path we estimate how the mean follow-up outcome would change if the abundance of the genus, on a log-ratio scale, were raised by one standard deviation in every participant, in total and split into an indirect part through the mediator and a direct part through all other pathways (Díaz & Hejazi, 2020). Every model adjusts for 35 baseline covariates (Appendix J), whose missing values are completed by K-MICE (Appendix G.1). Within each registered batch of routes, we control the false discovery rate over its candidate paths by the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) and, more strictly, the probability of any false discovery among the three effects by Holm's procedure (Holm, 1979).

The campaign recovers chains that agree with reported biology. Fifty-seven of the 64 routes returned estimates, covering all five pools and all twenty-four outcomes; of 18,012 candidate paths, 10,661 returned all three effects and 7,219 were refused, nearly all for lack of overlap. Twelve indirect effects pass Benjamini–Hochberg at 0.05 and one also passes Holm (Appendix J). Three of them, shown in Figure 2b, reproduce chains whose two links have both been reported: *Prevotella* through branched-chain amino acids to blood glucose (Pedersen et al., 2016; Wang et al., 2011); *Christensenellaceae* through visceral fat to liver elasticity, the path that passes Holm (Goodrich et al., 2014; Beaumont et al., 2016; van der Poorten et al., 2008); and *Akkermansia* through VLDL triglycerides to the ankle–brachial index (Dao et al., 2016; Aday et al., 2018).

6 Related work

Mediation identification. Natural effects are due to Robins & Greenland (1992) and Pearl (2001), and path-specific effects to Avin et al. (2005). The cross-world condition is untestable (Robins & Richardson, 2010); Imai et al. (2010) probe it by sensitivity analysis, whereas we report identification status separately. Our obstruction indicator flags the exposure-induced mediator–outcome confounding analysed by Tchetgen Tchetgen & VanderWeele (2014).

Screened structure discovery. Restricting search to candidate neighbours goes back to Sparse Candidate (Friedman et al., 1999) and MMHC (Tsamardinos et al., 2006); ours differs by always admitting the registered role pairs. NOTEARS (Zheng et al., 2018) and DAGMA (Bello et al., 2022) optimise a dense $p \times p$ matrix under an $O(p^3)$ acyclicity constraint and are not run at our scales.

Adjustment and estimation. Flexible nuisance estimation is standard (van der Laan et al., 2007; Chernozhukov et al., 2018); we add that one fixed channel-additive class suffices on the evaluated families. Complete criteria exist for sets valid throughout a Markov equivalence class (Perković et al., 2018), which a sampled family need not form, and Hu & van der Pas (2025) rank sets by posterior probability of validity; our adjustment set must instead pass equation 3 on every supporting graph or is withheld. Exposure-only predictors and near-instruments (Myers et al., 2011; Pearl, 2011) are admitted with their overlap cost reported. Li et al. (2025) introduced the decomposition we inherit; we replace what its three stages hand on, and the graph family only compiles one set.

7 Conclusion

CHARM estimates mediation effects end to end from incomplete, high-dimensional data. Its discovery scales to ten thousand variables. When the learned structures disagree, it never issues an invalid adjustment set. The estimator is accurate on both the linear and nonlinear mechanisms we tested. On the Human Phenotype Project, the chains it recovers survive multiple-testing correction and reproduce reported biology.

ETHICS STATEMENT

The real data come from the Human Phenotype Project (Shilo et al., 2021; Reicher et al., 2025), and we analysed them on the project's research platform with access granted by the project. All other experiments use simulated data.

REPRODUCIBILITY STATEMENT

Appendices A and B give the identification conditions and the proof of Proposition 1, with the compiler in execution order in Algorithm 1; Appendices C and D specify the effect stage and Appendix G the discovery stage with the settings of every reported run. For each of Tables 1–3, Appendix F gives the data-generating process, the comparison configurations and the metric definitions, together with the structural seeds of Tables 1 and 3, and Appendix I reports every replicate of the end-to-end campaign. The settings of the bridge and of the effect stage were fixed before any effect estimate was examined. For the HPP campaign of Section 5, Appendix J gives the cohort, the registered plan, the variables, the settings and every significant estimate.

AI USE STATEMENT

We used generative AI tools in the execution of this research: to give feedback on the methodology and the experimental design, to implement methods and data-processing code, and to help interpret results. We did not use them to generate synthetic data sets, which come from the programmatic generators of Appendix F, or to formulate or prove mathematical claims; qualitative or thematic data analysis is not applicable to this work. We also used them to search for and summarise related literature, to draft parts of the paper, to edit it for readability, to translate drafts for the authors' review and to write figure scripts. The authors reviewed all AI-assisted work: code was tested, reported numbers were checked against the recorded results, references were checked against the published sources, and all text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Aaron W. Aday, Patrick R. Lawler, Nancy R. Cook, et al. Lipoprotein particle profiles, standard lipids, and peripheral artery disease incidence. *Circulation*, 138(21):2330–2341, 2018.

Bryan Andrews, Joseph Ramsey, and Gregory F. Cooper. Learning high-dimensional directed acyclic graphs with mixed data-types. In *Proceedings of the 2019 ACM SIGKDD Workshop on Causal Discovery*, volume 104 of *Proceedings of Machine Learning Research*, pp. 4–21, 2019.

Bryan Andrews, Joseph Ramsey, Ruben Sanchez-Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of DAGs using the best order score search and grow-shrink trees. In *Advances in Neural Information Processing Systems 36*, 2023.

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 357–363, 2005.

Michelle Beaumont, Julia K. Goodrich, Matthew A. Jackson, et al. Heritable components of the human fecal microbiome are associated with visceral fat. *Genome Biology*, 17:189, 2016.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems 35*, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Hengrui Cai, Rui Song, and Wenbin Lu. ANOCE: Analysis of causal effects with multiple mediators via constrained structural learning. In *International Conference on Learning Representations*, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014.

Hoang Dang, Luan Pham, and Minh Nguyen. Effect-level validation for causal discovery. *arXiv preprint arXiv:2602.08340*, 2026.

Maria Carlota Dao, Amandine Everard, Judith Aron-Wisnewsky, et al. Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut*, 65(3):426–436, 2016.

Iván Díaz and Nima S. Hejazi. Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):661–683, 2020.

Chris Chinenye Emezue, Alexandre Drouin, Tristan Deleu, Stefan Bauer, and Yoshua Bengio. Benchmarking Bayesian causal discovery methods for downstream treatment effect estimation. In *ICML Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

Helmut Farbmacher, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler. Causal mediation analysis with double machine learning. *The Econometrics Journal*, 25(2):277–300, 2022.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021. doi: 10.3982/ECTA16901.

Nir Friedman, Iftach Nachman, and Dana Pe'er. Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 206–215, 1999.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Julia K. Goodrich, Jillian L. Waters, Angela C. Poole, et al. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, 2014.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Zhongyi Hu and Stéphanie van der Pas. Selecting valid adjustment sets with uncertain causal graphs. *arXiv preprint arXiv:2511.01662*, 2025.

Martin Huber, Yu-Chin Hsu, Ying-Ying Lee, and Layal Lettry. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 35(7):814–840, 2020.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

Jeremiah Jones, Ashkan Ertefaie, and Robert L. Strawderman. Causal mediation analysis: selection with asymptotically valid inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):678–700, 2025.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

Yulong Li, Xiwei Liu, Feilong Tang, Ming Hu, Jionglong Su, Zongyuan Ge, Imran Razzak, and Eran Segal. Decoding causal structure: End-to-end mediation pathways inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Romain Lopez, Jan-Christian Hütter, Jonathan K. Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. In *Advances in Neural Information Processing Systems*, 2022.

Jessica A. Myers, Jeremy A. Rassen, Joshua J. Gagne, Krista F. Huybrechts, Sebastian Schneeweiss, Kenneth J. Rothman, Marshall M. Joffe, and Robert J. Glynn. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11): 1213–1222, 2011. doi: 10.1093/aje/kwr364.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 411–420, 2001.

Judea Pearl. Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, 174(11):1223–1227, 2011. doi: 10.1093/aje/kwr352.

Helle Krogh Pedersen, Valborg Gudmundsdottir, Henrik Bjørn Nielsen, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, 535:376–381, 2016.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017. doi: 10.1007/s41060-016-0032-z.

Joseph D. Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A. Barnes, and Clark Glymour. TETRAD—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

Lee Reicher, Smadar Shilo, Anastasia Godneva, et al. Deep phenotyping of health–disease continuum in the Human Phenotype Project. *Nature Medicine*, 31:3191–3203, 2025.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9-12):1393–1512, 1986.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

James M. Robins and Thomas S. Richardson. Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, pp. 103–158. Oxford University Press, 2010.

Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

scikit-learn developers. sklearn.impute.IterativeImputer. `https://scikit-learn.org/1.4/modules/generated/sklearn.impute.IterativeImputer.html`, 2024. scikit-learn 1.4.2 documentation. Accessed 2026-09-10.

Smadar Shilo, Noam Bar, Ayya Keshet, et al. 10 K: a large-scale prospective longitudinal study in Israel. *European Journal of Epidemiology*, 36:1187–1194, 2021.

Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, 25(2):282–291, 2014.

Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):25, 2007. doi: 10.2202/1544-6115.1309.

David van der Poorten, Kerry-Lee Milner, Jason Hui, et al. Visceral fat: a key mediator of steatohepatitis in metabolic liver disease. *Hepatology*, 48(2):449–457, 2008.

Thomas J. Wang, Martin G. Larson, Ramachandran S. Vasan, et al. Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, 17:448–453, 2011.

Marcel Wienöbst, Max Bannach, and Maciej Liśkiewicz. Extendability of causal graphical models: Algorithms and computational complexity. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1248–1257, 2021.

Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition, 2017.

Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, Pantel Vokonas, Lihui Zhao, Jinchi Lv, Andrea Baccarelli, Lifang Hou, and Lei Liu. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9492–9503, 2018.

A ESTIMANDS AND IDENTIFICATION CONDITIONS

Estimand convention. With $M(x)$ the potential mediator and $Y(x, m)$ the potential outcome, the four nested means of equation 1 are $\eta_{ab} = \mathbb{E}[Y\{x_a, M(x_b)\}]$ for $a, b \in \{0, 1\}$. This paper uses the pure natural direct and total natural indirect pairing throughout, $\mathrm{NDE} = \eta_{10} - \eta_{00}$ and $\mathrm{NIE} = \eta_{11} - \eta_{10}$, which sum to $\mathrm{TE} = \eta_{11} - \eta_{00}$ with no residual term. The remaining mean η_{01} belongs to the total-direct and pure-indirect pairing. It is computed and stored as a diagnostic, never enters the three reported contrasts, and never gates them. Under an $X \times M$ interaction the two pairings differ, so a decomposition that mixes them does not sum to the total effect.

Maintained assumptions. Identification is argued with respect to a fixed adjustment set C. For all $x, x' \in \{x_0, x_1\}$ and all m in the support of M, the argument invokes the following.

$$
\begin{aligned}
&\text{(A1) invariance} && C(x) = C;\\
&\text{(A2) consistency} && X = x \Rightarrow M = M(x), \quad (X, M) = (x, m) \Rightarrow Y = Y(x, m),\\
& && Y(x) = Y\{x, M(x)\};\\
&\text{(A3) single world} && M(x) \perp X \mid C, \quad Y(x, m) \perp X \mid C, \quad Y(x, m) \perp M \mid X, C;\\
&\text{(A4) cross world} && Y(x, m) \perp M(x') \mid C;\\
&\text{(A5) positivity} && P(X = x \mid C = c) > 0,\\
& && \mathrm{P}_{M \mid X = x', C = c} \ll \mathrm{P}_{M \mid X = x, C = c}
\end{aligned}
\tag{7}
$$

for almost every c. With a continuous exposure the first condition is read as a conditional density of X given $C = c$ that is positive on a neighbourhood of each target level, and the conditional laws entering equation 8 are taken in versions continuous in x at x_0 and x_1, so that conditioning on the null event $X = x_a$ does not depend on the version chosen. The second condition is read through a common dominating measure for the mediator, counting measure for a discrete M and Lebesgue measure for a continuous one, and requires the law of M given (x', c) to be absolutely continuous with respect to its law given (x, c), so that $\mu_Y(x_a, \cdot, c)$ is defined wherever the mediator law of the other arm puts mass.

Identification. Under (A1) to (A5) the nested means are identified from the observed-data law by the mediation g-formula (Robins & Greenland, 1992; Pearl, 2001). Writing $\mu_Y(x, m, c) = \mathbb{E}[Y \mid X = x, M = m, C = c]$ for the outcome mean and $F_M(m \mid x, c) = P(M \le m \mid X = x, C = c)$ for the mediator law,

$$
\begin{aligned}
\eta_{ab} &= \mathbb{E}_C\big[\mathbb{E}\{Y(x_a, M(x_b)) \mid C\}\big]\\
&= \mathbb{E}_C\Big[\int \mathbb{E}\{Y(x_a, m) \mid M(x_b) = m, C\}\, dF_{M(x_b)}(m \mid C)\Big]\\
&= \mathbb{E}_C\Big[\int \mathbb{E}\{Y(x_a, m) \mid C\}\, dF_{M(x_b)}(m \mid C)\Big]\\
&= \mathbb{E}_C\Big[\int \mu_Y(x_a, m, C)\, dF_M(m \mid x_b, C)\Big].
\end{aligned}
\tag{8}
$$

The second line is iterated expectation over $M(x_b)$. The third line drops the conditioning on $M(x_b)$ and is the only step that uses cross-world exchangeability (A4). The fourth replaces both potential quantities by observed ones, using single-world exchangeability (A3) for $Y(x_a, m)$ given X and given M, and for $M(x_b)$ given X, together with consistency (A2) on the events where each replacement is made. Invariance (A1) keeps C the same object in both arms, and positivity (A5) makes every conditional quantity well defined, with the inner integral running over a mediator support that the outcome arm also reaches.

What a graph can and cannot supply. The three d-separation conditions of equation 3 are single-world back-door statements. On a graph that is accepted as causal and satisfies causal sufficiency and the causal Markov condition, they can support (A3) for the set they are evaluated on. They say nothing about (A4), which relates two incompatible exposure worlds and is not testable from observational data (Robins & Richardson, 2010), and no graphical check in this paper supplies it. Three judgements therefore remain distinct and are reported as separate status fields. Whether the functional in equation 8 can be evaluated on the estimation sample is numerical, and is settled by

the estimator and the integration rule. Whether a fixed C satisfies equation 3 on a supplied graph is graphical, and is settled relative to that graph alone. Whether the resulting number may be read as a natural effect is causal, and neither of the first two settles it. The argument above is a statement about the full-data law, and the missing-data conditions that connect it to an incomplete record are in Section 3.5. An estimate under fold-wise completion evaluates the same functional under the completed law, and maintaining (A1) to (A5) does not turn that number into the full-data natural effect.

B THE IDENTIFICATION-AWARE BRIDGE AND THE COMMON-SET COMPILER

This appendix gives the objects behind equation 3 and equation 11, proves the proposition stated in Section 3.4, and specifies the compiler, its diagnostics and its comparison configurations. Throughout, the assumptions are those of Appendix A and the numbering (A1) to (A5) is used without restatement.

For the graphs supporting π, the existence of a valid adjustment set for each graph does not imply the existence of one set valid across them.

$$\forall s\ \exists C_s \qquad \not\Longrightarrow \qquad \exists C\ \forall s. \tag{9}$$

B.1 CERTIFICATION OBJECTS

Fix a path $\pi = (X, M, Y)$. The candidate universe is $\mathcal{W}_\pi = \mathcal{V}_{\text{obs}} \setminus \{X, M, Y\}$, with $\mathcal{V}_{\text{obs}}$ the variables recorded in the study; the compiler of Section 3.4 draws its candidates from the smaller set $\mathcal{U}_\pi \subseteq \mathcal{W}_\pi$ that also excludes every registered candidate mediator. A set $C \subseteq \mathcal{W}_\pi$ is operationally eligible on a supplied graph G, written $\text{Elig}^{\text{op}}_G(C; \pi) = 1$, when its members are recorded variables outside $\{X, M, Y\}$ and none of them descends from X in G. Operational certification $\text{Cert}^{\text{op}}_G(C; \pi)$ adds to eligibility the three d-separation conditions of equation 3, evaluated in the mutilated graphs $G_{\underline{X}}$ and $G_{\underline{M}}$. The three are distinct single-world back-door problems and are not interchangeable, and Appendix A records what they can and cannot support. The superscript matters. A value of zero means uncertified on the graph supplied, and never that no valid set exists.

Throughout, G is a DAG on $\mathcal{V}_G$ containing $X \to M \to Y$ and $G_{\underline{Z}}$ is G with every edge leaving Z deleted. Write

$$B_{\text{GT}}(G, \pi) = \mathcal{V}_G \setminus \big(\text{De}_G(X) \cup \{X, M, Y\}\big),$$
$$\mathcal{L}_\pi(G) = \big(\text{De}_G(X) \setminus \{X, M, Y\}\big) \cap \text{An}_G(M) \cap \text{An}_{G\setminus\{M\}}(Y),$$

where $\text{De}_G(X)$ includes X. Every vertex therefore lies in exactly one of $\{X, M, Y\}$, $\text{De}_G(X) \setminus \{X, M, Y\}$, or B_{GT}. The set B_{GT} is a theoretical benchmark defined on a fully observed true graph. It is neither the compiled set $\widehat{C}_\pi$ of the compiler specified below, which is capped at K and read off learned structure, nor the design-supplied set C_{fix} of Appendix B.6, which is registered by the study design and never derived from a graph.

B.2 THE ORACLE NON-DESCENDANT SET

Proposition 1 (Oracle non-descendant set)**.** *$B_{\text{GT}}(G, \pi)$ satisfies the three d-separation conditions of equation 3 if and only if $\mathcal{L}_\pi(G) = \varnothing$.*

Proof. **Two auxiliary facts.** (a) If $A \in \text{Pa}_G(X)$ then $A \in B_{\text{GT}}$. Indeed $A \notin \text{De}_G(X)$, since $A \to X$ together with a directed path $X \rightsquigarrow A$ would close a cycle; and $A \notin \{X, M, Y\}$, since $A = M$ gives the cycle $X \to M \to X$ and $A = Y$ the cycle $X \to M \to Y \to X$. (b) No collider lying in $\text{De}_G(X) \setminus \{X\}$ is opened by conditioning on $\{X\} \cup B_{\text{GT}}$: a descendant of such a collider that lay in B_{GT} would itself be a descendant of X, contradicting the definition of B_{GT}, and a descendant equal to X would close a cycle.

Sufficiency. Assume $\mathcal{L}_\pi(G) = \varnothing$. In $G_{\underline{X}}$ the exposure retains only incoming edges, so any path out of X begins $X \leftarrow A$ with $A \in \text{Pa}_G(X)$. That A is a non-collider on the path, because it emits an edge into X, and by (a) it lies in B_{GT}; conditioning on B_{GT} therefore blocks the path at its first interior vertex. This holds for every path from X to M and every path from X to Y, giving the first two conditions unconditionally, and they do not use the hypothesis on $\mathcal{L}_\pi$.

For the third, suppose towards a contradiction that some path p between M and Y is open given $\{X\} \cup B_{\mathrm{GT}}$ in $G_{\underline{M}}$. Every non-collider of p must lie outside $\{X\} \cup B_{\mathrm{GT}}$, hence in $\mathrm{De}_G(X) \setminus \{X, M, Y\}$. We claim p carries no open collider. By (b) a collider inside $\mathrm{De}_G(X) \setminus \{X\}$ is closed, so an open collider c would have to lie in $\{X\} \cup B_{\mathrm{GT}}$. Suppose first $c \in B_{\mathrm{GT}}$ and let a be a neighbour of c on p, so that $a \to c$. Two adjacent vertices of a path cannot both be colliders, since the edge between them would have to point into each; so a is a non-collider or an endpoint. If a is a non-collider then $a \in \mathrm{De}_G(X)$ by the preceding sentence, whence $c \in \mathrm{De}_G(X)$, contradicting $c \in B_{\mathrm{GT}}$. If $a = M$ the edge $M \to c$ is absent from $G_{\underline{M}}$, and if $a = Y$ then $c \in \mathrm{De}_G(Y) \subseteq \mathrm{De}_G(X)$, again a contradiction. Suppose instead $c = X$. Both neighbours of X on p then lie in $\mathrm{Pa}_G(X) \subseteq B_{\mathrm{GT}}$ by (a); each is a non-collider by the adjacency argument, so each is conditioned on and blocks p, unless it is an endpoint, and $M \to X$ is absent from $G_{\underline{M}}$ while $Y \to X$ would close a cycle. Hence p has no collider at all and its interior lies in $\mathrm{De}_G(X) \setminus \{X, M, Y\}$. In $G_{\underline{M}}$ the mediator has no outgoing edges, so p leaves M as $M \leftarrow \cdot$; a collider-free path with that orientation is a fork, with an apex L from which directed paths run to M and to Y. The second of those avoids M, since p meets M only at its endpoint. Hence $L \in \mathrm{De}_G(X) \setminus \{X, M, Y\}$, $L \in \mathrm{An}_G(M)$ and $L \in \mathrm{An}_{G \setminus \{M\}}(Y)$, that is $L \in \mathcal{L}_\pi(G)$, contradicting the hypothesis.

Necessity. Assume $v \in \mathcal{L}_\pi(G)$ and fix directed paths $v \rightsquigarrow M$ and $v \rightsquigarrow Y$, the latter avoiding M. The two may share vertices beyond v, so let L'' be their *last* common vertex; a naive concatenation at v would in general yield a walk rather than a path. Being on a directed path out of v, L'' again lies in $\mathrm{De}_G(X) \setminus \{X, M, Y\}$ and is an ancestor of M and, avoiding M, of Y, so $L'' \in \mathcal{L}_\pi(G)$. Beyond L'' the two continuations are vertex-disjoint, so concatenating them gives a simple path $M \Leftarrow L'' \Rightarrow Y$ whose interior lies in $\mathrm{De}_G(X)$. None of its edges leaves M, so it survives in $G_{\underline{M}}$; L'' and every interior vertex is a non-collider; and no eligible set may contain a descendant of X, so none of them is conditioned on. The path is therefore open given $\{X\} \cup C$ for *every* eligible C, and in particular for B_{GT}, so the third condition fails. □

The necessity direction proves more than the proposition states. A non-empty $\mathcal{L}_\pi(G)$ defeats the third condition for *every* eligible C, not only for B_{GT}, and that stronger statement is the one used to define the obstruction indicator in the panel diagnostics below. The sufficiency argument also shows that the first two conditions hold for B_{GT} unconditionally, without any hypothesis on $\mathcal{L}_\pi$, so the whole content of the proposition sits in the mediator–outcome condition.

B.3 Exposure exogeneity does not settle the third gate

An exogenous exposure disposes of the first two conditions and leaves the third untouched. Take $\mathcal{V}_G = \{X, L, M, Y\}$ with edges $X \to L$, $X \to M$, $X \to Y$, $L \to M$, $L \to Y$ and $M \to Y$. Here X is a root, so the first two conditions hold with $C = \varnothing$. However $L \in \mathcal{L}_\pi(G)$, and $B_{\mathrm{GT}} = \varnothing$ because L descends from X. The fork $M \leftarrow L \to Y$ is open given X in $G_{\underline{M}}$, and by the necessity argument no eligible set closes it. A vertex in the causal future of the exposure that also influences mediator and outcome is exposure-induced mediator–outcome confounding (Tchetgen Tchetgen & VanderWeele, 2014). Blocking the route it opens would require conditioning on a descendant of X, which eligibility forbids, so the obstruction is a property of the graph rather than a failure of the search. Equation 3 keeps the third condition separate for exactly this reason.

B.4 The compiler

The compiler draws candidates from $\mathcal{U}_\pi := \mathcal{V}_{\mathrm{obs}} \setminus (\{X, Y\} \cup \mathcal{M})$, where $\mathcal{V}_{\mathrm{obs}}$ contains the variables recorded in the study and $\mathcal{M}$ contains every registered candidate mediator. Algorithm 1 lists the steps in execution order, and the paragraphs below define each.

The compiler of Section 3.4 reads node evidence from the whole family $\mathbb{G}^{\mathrm{all}} = (G_s)_{s \in \mathcal{I}_{\mathrm{all}}}$ of S draws, including the draws that do not contain both path edges. Restricting the evidence to the supporting indices $\mathcal{I}_\pi$ would condition it on graph-level realisation of the same two edges whose

Algorithm 1 Common-set compiler, in execution order

Require: family $(G_s)_{s=1}^{S}$, $S = 30$; path $\pi = (X, M, Y)$; registered mediators $\mathcal{M}$; post-repair auxiliary weights w with operating threshold τ_{aux}; $Q = 12$, $K = 2$; panels $\mathcal{I}^A = \{1, \dots, 15\}$, $\mathcal{I}^B = \{16, \dots, 30\}$.

Ensure: $\widehat{C}_\pi$, or $(\perp, \Xi_\perp)$ with every applicable reason; in both cases the diagnostics of C_π^{prop}.

1: Compute $(N_{\text{anc}}, N_{\text{pa}}, N_{\text{de}})$ by equation 10 on all S graphs, then $\mathcal{U}_\pi^0$ and $\mathcal{D}_\pi = \text{Top}_Q(\mathcal{U}_\pi^0)$ by equation 11.
2: $\mathcal{C} \leftarrow \{C \subseteq \mathcal{D}_\pi : 1 \le |C| \le K\}$ if $\mathcal{D}_\pi \neq \varnothing$, else $\mathcal{C} \leftarrow \{\varnothing\}$.
3: **for** $C \in \mathcal{C}$, $P \in \{A, B\}$ **do** $n^P(C) \leftarrow \sum_{s \in \mathcal{I}_\pi^P} \text{Cert}_{G_s}^{\text{op}}(C; \pi)$, $r^P(C) \leftarrow \rho\big(n^P(C), m_\pi^P\big)$.
4: $C^{\text{pool}} \leftarrow$ first element of $\mathcal{C}$ under equation 12, every rate formed by ρ.
5: **for** $v \in \mathcal{V}_{\text{obs}} \setminus (\{X, M, Y\} \cup \mathcal{M})$, $R \in \{X, M, Y\}$ **do** $q_R(v) \leftarrow \max\{|w_{vR}|, |w_{Rv}|\}/\tau_{\text{aux}}$.
6: $E(v) \leftarrow \mathbf{1}\big\{\min_R q_R(v) \ge 0.25,\ \sum_R \mathbf{1}[q_R(v) \ge 1] \ge 2\big\}$; $(e_1, e_2, \dots) \leftarrow \{v : E(v) = 1\}$ sorted by $o(v)$.
7: $\omega \leftarrow \mathbf{1}\big\{\min_{v \in C^{\text{pool}}} N_{\text{anc}}(v) < 0.25S\big\} \vee \mathbf{1}\big\{n^A(C^{\text{pool}}) = 0 \vee n^B(C^{\text{pool}}) = 0\big\}$, with $\min_\varnothing = 0$.
8: $C_\pi^{\text{prop}} \leftarrow \begin{cases} \{e_1, e_2\}, & \omega = 1,\ e_2 \text{ exists}, \\ \varnothing, & \omega = 1,\ e_2 \text{ does not exist}, \\ C^{\text{pool}}, & \omega = 0. \end{cases}$
9: $t_{\text{fb}} \leftarrow \mathbf{1}\big\{\omega = 1,\ e_3 \text{ exists},\ o_{1:3}(e_2) = o_{1:3}(e_3)\big\}$; $t_{\text{pool}} \leftarrow 0$.
10: $T \leftarrow \{C \in \mathcal{C} : \sigma(C) = \sigma(C^{\text{pool}})\}$; $T_H \leftarrow \{C \in T : |C| = K,\ E(v) = 1\ \forall v \in C\}$.
11: **if** $|T| > 1$ **then**
12: **if** $|T_H| = 1$ **then** $C_\pi^{\text{prop}} \leftarrow$ the element of T_H **else** $t_{\text{pool}} \leftarrow 1$.
13: **end if**
14: **for** $s \in \mathcal{I}_\pi^A \cup \mathcal{I}_\pi^B$ **do** $c_s \leftarrow \text{Cert}_{G_s}^{\text{op}}(C_\pi^{\text{prop}}; \pi)$, $z_s \leftarrow \mathbf{1}\{\mathcal{L}_\pi(G_s) \neq \varnothing\}$.
15: $\Xi_\perp \leftarrow$ the true conditions among $C_\pi^{\text{prop}} = \varnothing$, $t_{\text{fb}} = 1$, $t_{\text{pool}} = 1$, $\min_P m_\pi^P < 3$, $\exists s : z_s = 1$, $\exists s : c_s = 0$.
16: **return** $\widehat{C}_\pi = C_\pi^{\text{prop}}$ if $\Xi_\perp = \varnothing$, or if $\omega = 0$ and $t_{\text{pool}} = 1$ is the only condition in $\Xi_\perp$; else $(\perp, \Xi_\perp)$.

$\mathcal{I}_\pi^P$: graphs of panel P that support π, of size m_π^P. $\rho(a, b) = a/b$ for $b > 0$ and 0 otherwise; it enters ranking and the tie comparison only, and a diagnostic with a zero denominator is reported as not assessed. $r^{\text{pool}} = \rho(n^A + n^B, m_\pi^A + m_\pi^B)$, $\bar{r} = (r^A + r^B)/2$. $o(v) = \big(-\min_R q_R,\ -\bar{q}^{\,\text{geo}},\ -\#\{R : q_R \ge 1\},\ \text{idx}\big)(v)$, with $\bar{q}^{\,\text{geo}}$ the geometric mean of the three ratios, 0 if any is 0, and $o_{1:3}$ its first three keys. $\sigma(C) = \big(r^{\text{pool}}, \min_P r^P, \bar{r}, \mathbf{1}[C \neq \varnothing], \{\{(N_{\text{anc}}, N_{\text{pa}}, N_{\text{de}})(v) : v \in C\}\}\big)(C)$, with $\{\{\cdot\}\}$ an unordered multiset. The record keeps $n^P(C_\pi^{\text{prop}})$, m_π^P and every s with $z_s = 1$.

recurrence defines the path-selection score. For $v \in \mathcal{U}_\pi$,

$$N_{\text{anc}}(v) = \sum_{s \in \mathcal{I}_{\text{all}}} \mathbf{1}\Big\{\textstyle\sum_{R \in \{X,M,Y\}} \mathbf{1}[v \in \text{An}_{G_s}(R)] \ge 2,\ v \notin \text{De}_{G_s}(X)\Big\},$$
$$N_{\text{pa}}(v) = \sum_{s \in \mathcal{I}_{\text{all}}} \sum_{R \in \{X,M,Y\}} \mathbf{1}\{v \in \text{Pa}_{G_s}(R)\}, \qquad N_{\text{de}}(v) = \sum_{s \in \mathcal{I}_{\text{all}}} \mathbf{1}\{v \in \text{De}_{G_s}(X)\}. \tag{10}$$

The three counts live on different scales by construction. $N_{\text{anc}}(v)$ is an indicator per draw, so $N_{\text{anc}}(v) \le S$. $N_{\text{pa}}(v)$ accumulates once for each path role a vertex parents within a draw, so a vertex that is a parent of both M and Y in one graph contributes two for that draw and $N_{\text{pa}}(v) \le 3S$. $N_{\text{de}}(v)$ counts draws in which v descends from X, with $\text{De}_{G_s}(X)$ including X itself.

The candidate domain. Using the counts of equation 10, the evidence core and the capped candidate domain are

$$\mathcal{U}_\pi^0 = \{v \in \mathcal{U}_\pi : N_{\text{anc}}(v) + N_{\text{pa}}(v) > 0,\ N_{\text{de}}(v) = 0\}, \qquad \mathcal{D}_\pi = \text{Top}_Q(\mathcal{U}_\pi^0). \tag{11}$$

The universe $\mathcal{U}_\pi$ already excludes every registered candidate mediator, so none can enter the set. A single draw placing v downstream of the exposure removes v. On $\mathcal{U}_\pi^0$ the vertices are ordered by the key $\big(-N_{\text{anc}}(v), -N_{\text{pa}}(v), \text{name}_{\text{can}}(v)\big)$, with name_{can} the canonical variable name. The order is

total and deterministic. In equation 11, Top_Q takes the first $Q = 12$ vertices in this order, and the position of a vertex is its evidence rank. Any core vertices cut by the cap are recorded.

Pooled candidate selection. The set is capped at $K = 2$. The compiler scores every nonempty subset of $\mathcal{D}_\pi$ with at most K members rather than taking a prefix of the evidence order, so a domain of twelve vertices gives $12 + 66 = 78$ candidates, and an empty domain yields only the empty candidate, which is never issued. Each candidate C is evaluated by equation 3 on both fixed panels, giving counts $n^P(C)$ over the m_π^P graphs of panel P that support π. Candidates are ordered by

$$\Big(-\tfrac{n^A(C)+n^B(C)}{m_\pi^A+m_\pi^B},\ \ -\min\big\{\tfrac{n^A(C)}{m_\pi^A}, \tfrac{n^B(C)}{m_\pi^B}\big\},\ \ \textstyle\sum_{v\in C}\mathrm{rank}(v),\ \ \mathrm{key}_{\mathrm{can}}(C)\Big), \tag{12}$$

and the first candidate is the pooled winner. Writing $r^P(C) = n^P(C)/m_\pi^P$, a panel rate or the pooled rate with a zero denominator is taken as 0 in this order and in the tie comparison below, and the mean panel rate $\bar{r}(C) = \{r^A(C) + r^B(C)\}/2$ enters that comparison but not the order. The pooled rate leads because it uses every supporting graph. The minimum panel rate breaks ties against a candidate that certifies in one half of the family and nowhere in the other. The evidence-rank sum then prefers the structurally better attested candidate and the canonical key makes the listing total, but neither is treated as evidence by the ambiguity check below. Nothing in equation 12 reads the true graph, the effect estimates or the path-selection score.

Two conditions withdraw the winner. The first is low support. Each member of the pooled winner carries a common-ancestor support count from the domain construction, and the condition fires when the smallest of these counts falls below a registered fraction of S, fixed at 0.25, an empty winner counting as 0. The second is a one-sided panel. It fires when the winner records zero certifications within panel A or within panel B, so that one half of the family never certifies the candidate the pooled rate selected. The two conditions are independent and either alone is sufficient. A pooled winner can therefore be withdrawn while its pooled rate is high, which is the intended behaviour when that rate is carried entirely by one panel.

The fallback proposal. When a condition fires the compiler looks outside the evidence order. For every vertex other than the three path roles and the publicly registered mediator candidates, it forms three role ratios, one for each of X, M and Y, as the largest absolute weight between the vertex and that role in either direction in the post-repair auxiliary weights of Appendix G.3, divided by their fixed operating threshold. A vertex is eligible for the fallback when its smallest role ratio is at least 0.25 and at least two of its three ratios reach 1. Eligible vertices are ordered by decreasing minimum role ratio, then decreasing geometric mean of the three, taken as 0 when any ratio is 0, then decreasing count of ratios reaching 1, then increasing canonical index, and the first two become the adjustment set. If the second and third eligible vertices agree on the first three keys, the canonical index alone would decide membership, and no set is issued. The fallback ignores orientation by construction, since a ratio takes the larger of the two directed weights, which is what makes it usable where orientation is the weakest part of the discovery output.

The fallback is not drawn from $\mathcal{U}_\pi^0$ and does not consult equation 10. It ranks on auxiliary edge strength alone and carries no construction-time guarantee of freedom from exposure descendants. A fallback proposal is issued only after the acceptance rule below certifies it on every supporting graph, and certification includes eligibility in each of them. Freedom from descendants in the true graph is not implied and is measured against the revealed graph per run.

Ties, acceptance and refusal. Before a proposal is accepted, every candidate that matches the provisional winner on exact pooled rate, minimum panel rate, mean panel rate and nonemptiness, and whose members carry the same unordered triples of common-ancestor support, direct-parent roles and exposure-descendant count, counts as tied with it; canonical names and ordinal evidence ranks are not treated as evidence. When more than one distinct candidate is tied, the tie is resolved only if exactly one tied candidate of size K has only fallback-eligible members, and otherwise it is a reason for refusal. A support tie at the boundary of the candidate domain is recorded but does not by itself cause refusal. The tie check always runs on the pooled winner, also when the fallback supplied the proposal, and a resolved pooled tie replaces the proposal. A fallback tie flagged before that resolution is kept, so the path is still refused when the fallback route had a tie. A proposal from either route is accepted only when it is nonempty, certifies under equation 3 on every graph that supports π in panel A and in panel B, and each panel holds at least three supporting graphs,

the last condition excluding a vacuous pass on a panel with no support. A supporting graph with a non-empty obstruction set $\mathcal{L}_\pi$ also leads to refusal. When an unresolved pooled tie is the only reason and the fallback was not used, the pooled winner already meets every acceptance condition and is issued as a representative of the tied candidates, with the tie recorded as a diagnostic. When the bridge refuses, it issues no adjustment set, written $\perp$, rather than the empty set, the effect stage does not run for that path, and the run record of Appendix E lists every reason that applies together with the certification counts, supporting-graph counts and obstruction graphs of the proposal. The planned denominator keeps the path, which is recorded as a case with no issued set. The empty set $\varnothing$ is a different object, a deliberately specified unadjusted comparison set that the arm of Appendix B.6 estimates with.

A fixed rule and role filtering. The selection rule is fixed as a whole and recorded with every input it reads, and changing the evidence source, the descendant filter, either order, the cap or a threshold defines a different rule that requires a fresh run. The compiler applies no filter to exposure-only predictors, and the registered candidate mediators are the only variables it excludes by role. A vertex that mainly predicts the exposure can enter $\widehat{C}_\pi$ when the structural evidence and panel certification support it. The cost of such a member is reported through the overlap diagnostics of Appendix C rather than removed after the fact, and excluding such predictors would be a different selection rule requiring a fresh run.

Development and assessment data. The candidate rule, its two rescue conditions and the acceptance rule were fixed after development campaigns at $p = 500$ had been run and their graphs revealed, so the rule is fitted to that evidence and cannot be confirmed by it, and those campaigns are not reported. The paper reports a fresh assessment of the bridge on seeds disjoint from a registered list of earlier seeds, run under a plan fixed before data generation and with the truth revealed only after every run had finished (Appendix F, Section 4.3). The subset sizes and the representative issuance above were adopted after that assessment; replayed on its stored graph families, they reproduce every selection of all fifty replicates, and the stress tests of Appendix I were run with them from the start.

External pre-exposure check. An external pre-exposure check $\mathrm{PreAttest}_{\mathrm{ext}}(C;\pi)$ takes values in $\{\mathrm{pass}, \mathrm{fail}, \mathrm{not\ assessed}\}$. It never filters or reselects a fixed set, and a learned $N_{\mathrm{de}}(v) = 0$ never substitutes for it. No such check was run in the experiments reported here.

B.5 Panel diagnostics

The ensemble is partitioned into two fixed panels before any candidate is scored. With $\mathcal{I}_\pi^P = \mathcal{I}^P \cap \mathcal{I}_\pi$ the graphs of panel P that support π, and m_π^P its size,

$$r_\bullet^P(\pi) = \frac{1}{m_\pi^P} \sum_{s \in \mathcal{I}_\pi^P} Z_s^\bullet, \qquad \bullet \in \{\mathrm{obst}, \mathrm{cert}\},$$

where Z_s^{obst} indicates a non-empty obstruction set on G_s and Z_s^{cert} is its complement times $\mathrm{Cert}_{G_s}^{\mathrm{op}}(C_\pi^{\mathrm{prop}};\pi)$, with C_π^{prop} the final proposal of Algorithm 1. The rates therefore exist for a refused path as well as for an issued one, for which $\widehat{C}_\pi = C_\pi^{\mathrm{prop}}$. A rate is reported as not assessed when a panel supports the path nowhere. A graph that does not support the path lies outside these denominators, yet it still contributed to the node evidence of equation 10, so the denominator for the rates and the evidence base for the domain are deliberately different objects.

The panels are selection inputs and are reported afterwards on the final proposal. Both enter candidate scoring, the rescue decision and the final certification, so neither is held out, the reported rates are not an independent assessment of the proposal, and an A–B difference is neither generalisation error nor a probability that identification holds. No standard error accompanies these rates, because a binomial interval describes the sampling fluctuation of a fixed candidate on graphs that did not choose it, which is not the design in force here. Three prohibitions are enforced as checks rather than stated as advice. A member of the proposal is never removed on its own because a panel failed to support it, since the unit of selection is the whole proposal and a withdrawal replaces it through the fallback or leaves the path without a set. A candidate is never rescored after the fact under a

different panel assignment. A label calling a compiled set stable or weak never changes an estimate that is already defined.

B.6 Comparison configurations

Section 3.4 runs the compiler above. This subsection specifies the two arms against which that choice is compared and the diagnostic mode that produces no estimate. The arms are defined here; what they show is reported with the experiments, and nothing about their outcome is asserted in advance.

What the two arms isolate. Automating the adjustment set changes two things at once: a set is chosen rather than supplied, and a set is used rather than omitted. The arms separate them. The design-supplied arm holds everything fixed except the origin of C_π, so a difference between it and the compiled arm is attributable to the selection policy. The unadjusted arm removes adjustment altogether and serves as the practical no-adjustment comparator. Its outcome model also differs from that of the other two arms in one end-to-end replicate (Appendix I), so a difference between it and either of the other two reflects the absence of an adjustment set together with that change, and is not a single-factor contrast of the set.

The design-supplied arm. Here $C_\pi := C_{\mathrm{fix}}$ is registered by the study design and no compilation from the ensemble is performed. The following conditions, omitted from the main text, apply: C_{fix} enters every nuisance component as a whole and is exempt from the size limits of the compiler; it is never truncated to fit them; an empty C_{fix} must be registered in advance and is first reported as an unadjusted functional; a path for which the design registers no set is recorded as lacking a prerequisite rather than as vetoed by any graph; and because no compilation is performed, the whole ensemble is available for diagnostics. A simulation may additionally construct a variant whose set is valid on the true graph, and that variant alone may be described as oracle-valid; the arm as such carries no oracle status.

The unadjusted arm. Here $C_\pi := \varnothing$ by construction. The empty set is a deliberately specified comparison set, not a claim that no adjustment is needed on this generator, and it differs from $\perp$, the state in which the bridge issues no set and the path is not run. The arm is estimated normally and carries its own recorded reason, so a reader can tell it apart from a refusal in every table.

Comparison protocol. All arms consume the same rows, the same discovery and effect halves, the same split, the same training settings and the same integration rule, and differ in the origin of C_π. In the end-to-end experiment the compiled- and supplied-set arms also fitted the same outcome and mediator classes in every replicate, so a difference between them is attributable to the set; the unadjusted arm used a different outcome model in one replicate (Appendix I). Two limits apply. A difference in either direction is a statement about the data-generating processes examined and not about adjustment selection in general. Identical sets show that the compiler reproduces the design-supplied set on the evaluated processes. Set agreement alone does not establish validity, and identical estimates alone do not establish set agreement.

The existence diagnostic. A third configuration chooses no common set and produces no effect estimate. It reports, over a finite candidate class, how often some candidate is certified on a graph and whether the graphs that admit one admit the same one. Because it yields a different set per graph, those sets are never averaged into a single estimate. This configuration is defined for diagnostic use and was not run in the experiments reported here.

B.7 What the bridge does not prove

Five boundaries close this appendix. No theorem here shows that the compiler recovers B_{GT} or any other valid set, and Proposition 1 characterises the benchmark rather than the compiled output. The admissibility rates of Section 4.3 are measurements on one registered synthetic generator under a rule fixed in advance, and a high rate there is evidence about that generator rather than a guarantee carried to another one. The set size K is a registered constant and is not learned, so a benchmark

whose valid sets are larger than K places admissibility out of reach by construction. On the generator behind Section 4.3 every registered path carries a design-registered valid set of exactly two members, matching $K = 2$, so the reported rate describes that matched case and says nothing about the rule when valid sets and K differ in size. That reference set is a design object and not B_{GT}. Compatibility on sampled graphs is not validity on the true graph, since every rate in Appendix B is evaluated against members of an ensemble that no result here certifies as causal. Graphical certification of a set supports at most (A3), and only on a graph accepted as causal that also satisfies causal sufficiency and the causal Markov condition, while (A4) is untestable from observational data and is supplied by no check in this paper. The causal label on every path-level record is therefore withheld by default, and the numerical, graphical and causal statuses of Appendix A are reported as separate status fields.

C NUISANCE LEARNING, MISSINGNESS AND EXPOSURE SUPPORT

This appendix gives the cross-fitting scheme, the fixed learners with their training record, the bias identity behind the choice of flexible channel, the exposure-support diagnostics and the three missing-data modes of Section 3.5. Every quantity below is defined on the estimation sample $\mathcal{I}_{\mathrm{est}}$ and against the single adjustment set C_π delivered by Appendix B.

C.1 CROSS-FITTING FOLDS

The estimation sample is partitioned into K_{cf} outer folds $\{I_k\}$, and $I_{-k} = \mathcal{I}_{\mathrm{est}} \setminus I_k$ is the training complement of fold k. The count is written K_{cf} rather than K because the set-size cap of Appendix B is a different object and the two must not be read as one. Admissibility requires $2 \le K_{\mathrm{cf}} \le n_{\mathrm{est}}$ with every I_k and every I_{-k} non-empty. We set $K_{\mathrm{cf}} = 5$. In the end-to-end experiment the folds come from one dataset-level split plan, a shuffled partition seeded by a hash of the data-set identifier, the split purpose and the code version, so every path and every arm of a replicate uses the same folds. Without such a plan the partition is seeded by the global seed and the path identifier; the mechanism panel uses no shared plan, and its methods do not share folds (Appendix F). A path that cannot meet the condition, and for which no externally registered nuisance predictions exist, is recorded as not run with the reason that too few rows remain for a cross-fitted estimate. No point estimate is manufactured by training and validating on the same row.

Completion happens inside cross-fitting rather than before it. For each k the completion model and all of its tuning are fitted on I_{-k} alone and then kept fixed, and the fold keeps two objects, the completed training representation and the transform applied to I_k. The inputs used to predict an outcome on a validation row never include that row's own outcome, and the same exclusion holds for every other target-specific completion. These fold-specific completions are distinct from the discovery completion of Section 3.2, which is fitted on discovery rows and never reaches the effect stage. The option of Appendix G.1 that discards fitted models cannot be used here, because the held-out rows need the fitted transform.

The order in which columns are completed keeps a validation target out of its own predictors. The adjustment record is completed first, from the adjustment columns alone and with no path-role variable among the predictors, using the chained conditional updates of Appendix G.1, Bayesian ridge predictive draws and a registered cap of ten passes. This adjustment-block imputer operates on the raw scale without the standardization used at discovery, retains its fitted models for validation, and checks observed entries for exact equality on the raw scale. The exposure, the mediator and the outcome are then completed one at a time in that causal order, each from the columns already completed before it, so the outcome never predicts the mediator and no target predicts itself. These path-role completions use Bayesian ridge predictive draws for continuous targets and logistic class-probability draws for discrete targets; an empirical training distribution or a single-class fill covers the cases without sufficient predictors or class variation. Each target draws its own seed from the fold seed. A binary column is rounded and clipped to its two levels and a k-level categorical column to $\{0, \dots, k-1\}$, so a completed value lies in the registered support.

A completion stops with an error if any of five conditions fails. The training and validation rows must be disjoint and both non-empty, every registered column must be present, no training column may be entirely missing, the completed table must keep the shape of the raw one, and every completed

value must be finite. A sixth is checked afterwards: every observed cell must be returned unchanged, and a completion that alters one stops with an error.

Each fold-specific completion is stored together with the rows it was fitted on and the rows it was applied to, the columns it completed in order, its seed and its iteration cap, so the rows behind any prediction can be recovered and matched against the fold that produced it. This is what makes an out-of-fold claim checkable rather than asserted. What cross-fitting delivers is that separation of fitting from evaluation (Chernozhukov et al., 2018), and nothing about the validity of an interval follows from it.

C.2 The fixed effect architecture

The three channels of the base class equation 4 have fixed forms set before any effect estimate is examined. The exposure channel is linear, $f(x) = \beta_X x$, which keeps the direct effect clear of the attenuation a flexible fit in x would introduce. The mediator channel is a network in m alone and replaces the linear term $\beta_M m$ outright rather than sitting beside it as a penalised residual. The adjustment channel is a network over C_π. The base class and the mediator learner are fixed before outer-fold fitting and are not revisited inside a fold, and Appendix I records the classes fitted in the end-to-end experiment. One pre-registered diagnostic decides whether the declared interaction extension is added; the end-to-end experiment applies it to every path, and the effect-only panel of Table 2 does not apply it (Appendix F). It tests exposure–mediator interaction once per path on the effect-stage data before outer-fold fitting: the outcome and the terms x, m, xm, m^2, m^3, and for a continuous exposure x^2 and x^3, are residualised on C_π by two-fold cross-fitted networks, and the residualised product term is tested against the others by an F-test at level 0.01 with a partial R^2 floor of 10^{-4}. When the test fires, the outcome model gains a varying-coefficient term $x\, g_{XM}(m)$ with a network of the same form, the effect is labelled supported with a warning, and the direct and total effects are labelled diagnostic. The test reads every effect-stage row, so the out-of-fold separation of the cross-fitting scheme covers the fitting of model parameters but not this structural decision.

Each channel network has two hidden layers of width 64 with rectified-linear activations, dropout 0.05 after the first, and a scalar linear output. Training minimises squared error with AdamW at learning rate 10^{-3}, weight decay 10^{-5}, batch size 128 and a fixed 50 epochs, with no early stopping and no validation-driven schedule. Features and target are standardised on the training fold and inverted after prediction. The global seed is 42; when component-stable seeding is enabled, as in the end-to-end experiment, the seed of each nuisance component is derived from the dataset-level split plan together with the path, the fold and the channel, and the configuration is recorded with the run so that a different combination cannot be reported under the same record.

C.3 The mediator learner and the exposure propensity

The default mediator learner is described below, and the end-to-end experiment of Table 3 fits its continuous mediator by a linear conditional mean (Appendix F). A continuous mediator is fitted as a conditional mean $\nu_M(x, c)$ on the standardised mediator, and the residual distribution around that fit is retained for the integration of Appendix D. That is the location–residual route, and it carries the working condition that the residual law does not vary with (x, c). A discrete mediator is fitted as a proper probabilistic classifier, which induces $\widehat{F}_M$ directly and carries no such condition, and the route taken is recorded on the path rather than inferred from the estimand. Both use a three-hidden-layer network of widths 128, 256 and 128 with rectified-linear activations and dropout 0.1, trained with AdamW at learning rate 10^{-4}, weight decay 10^{-5}, batch size 128 and 50 epochs, under squared error for a continuous mediator, binary cross-entropy for a binary one and categorical cross-entropy for a categorical one. For a binary exposure the propensity uses that same network and schedule as a classifier over C_π by default, and an L2-penalised logistic regression can be registered in its place; a continuous exposure, as in the end-to-end experiment, has no propensity model.

C.4 Why the mediator channel is the flexible one

Outcome misspecification reaches the indirect effect through one inner product rather than through goodness of fit in general. Let $\mu_{Y,\mathcal{F}}$ be the limiting projection of the outcome regression onto the working class, and define the residual of that projection at the comparison level together with the

mediator shift the exposure induces,

$$r_1(m,c) := \mu_Y(x_1,m,c) - \mu_{Y,\mathcal{F}}(x_1,m,c),$$
$$\mathrm{d}\Delta(m,c) = \mathrm{d}F_{C_\pi}(c)\big\{\mathrm{d}F_M(m \mid x_1,c) - \mathrm{d}F_M(m \mid x_0,c)\big\}. \tag{13}$$

When the remaining nuisances are consistent and the integral of equation 5 is evaluated exactly, the plug-in indirect effect has limiting bias

$$\widehat{\mathrm{NIE}} - \mathrm{NIE} \longrightarrow -\int r_1(m,c)\,\mathrm{d}\Delta(m,c). \tag{14}$$

Three readings follow and they do not all point the same way. Misspecification is not automatically fatal, since a residual orthogonal to the mediator shift leaves the indirect effect untouched and a model can fit poorly while returning the right contrast. Once the inner product in equation 14 is non-zero the bias is of constant order and does not diminish as the sample grows, so no quantity of data repairs a mediator channel whose residual is aligned with the shift. And because the residual is evaluated at the comparison level x_1, what matters is the fit in the arm the contrast is taken against rather than the average fit across arms. Equation 14 motivates which channel is given flexibility and is never an objective. Nothing in the procedure estimates the integral, which involves the very regression whose error is in question. Under fold-wise completion the right-hand side is a discrepancy relative to the completed law rather than a bias in the natural effect.

C.5 Exposure support

Write R_i^X for the indicator that X_i and every entry of $C_{\pi,i}$ are raw-observed. In the first mode the propensity is fitted on raw rows with that indicator identically one. In the second it is fitted on the training rows with $R_i^X = 1$ and predicted on raw-observed adjustment records. In the third it is fitted on the completed representation. The present design estimates no observation propensity for the exposure, so a weighted path whose exposure carries raw missingness has its overlap summaries labelled observation-conditional rather than full-data. Substituting a completed exposure inside the weighting mode is not allowed, and a training fold lacking either arm is recorded as an absent arm or as an unavailable overlap diagnostic.

With a prespecified $0 < \epsilon_X < 1/2$ the empirical overlap is read from the unclipped out-of-fold propensity through the band $\epsilon_X \le \widehat{e}_i \le 1 - \epsilon_X$, and the path-level violation fraction $v_X(\pi)$ is the share of estimation rows falling outside it. A separately prespecified $0 < \epsilon_{\mathrm{diag}} < 1/2$ defines the clipped $\widehat{e}_i^{\mathrm{diag}} = \mathrm{clip}(\widehat{e}_i, \epsilon_{\mathrm{diag}}, 1 - \epsilon_{\mathrm{diag}})$, whose only purpose is to keep a boundary prediction from dividing the diagnostic summaries by zero. Writing A_i^X for the availability indicator of the active mode and $X_i^\dagger$ for its exposure, the per-arm diagnostic weights and effective sample sizes are

$$a_i^{(a)} = \frac{A_i^X \mathbf{1}\{X_i^\dagger = x_a\}}{\widehat{e}_i^{\mathrm{diag},a}}, \qquad \mathrm{ESS}_a = \frac{\big(\sum_{i\in\mathcal{I}_{\mathrm{est}}} a_i^{(a)}\big)^2}{\sum_{i\in\mathcal{I}_{\mathrm{est}}}\{a_i^{(a)}\}^2}, \qquad a \in \{0,1\}, \tag{15}$$

with $\widehat{e}_i^{\mathrm{diag},1} = \widehat{e}_i^{\mathrm{diag}}$ and $\widehat{e}_i^{\mathrm{diag},0} = 1 - \widehat{e}_i^{\mathrm{diag}}$. An arm whose diagnostic weights sum to zero is assigned $\mathrm{ESS}_a = 0$ and recorded as absent rather than evaluated as $0/0$.

A positive violation fraction, a large boundary share, a small effective sample size or a heavy weight tail marks the record support-sensitive or extrapolating. None of them deletes a functional that remains finite and uniquely defined, since a finite-sample threshold on an estimated propensity does not demonstrate that population positivity fails. The clipped propensity stabilises summaries and never replaces the raw report, which is reported alongside it. A target arm with no usable training observation is a different state, and the path is recorded as not run because the declared procedure never executed. A continuous exposure is given no pseudo-binary propensity. Support is read instead from the empirical percentile of each declared level x_0, x_1: a level between the 5th and 95th percentiles is supported, one in either tail is supported with a warning, one outside the observed range by at most half a standard deviation is fragile and one further out is off support, and the path takes the worse of the two labels. The overlap cost of a variable that mainly predicts the exposure is reported here rather than removed upstream, which is the trade-off a near-instrument presents (Myers et al., 2011; Pearl, 2011) and the reason the compiler of Appendix B applies no filter to exposure-only predictors.

C.6 OBSERVED-ROW WEIGHTING

The weighting mode is prespecified and carries missingness conditions of its own. Let $\mathcal{U}_i^Y = (X_i, M_i, C_{\pi,i}, Y_i)$ and $\mathcal{U}_i^M = (X_i, C_{\pi,i}, M_i)$ collect the raw entries each nuisance needs, and let R_i^T indicate that all of them are observed for row i. The observation-model predictor set $\mathcal{A}_R$ is fixed before fitting and every variable in it must be raw-observed on every analysis row, which forbids predicting whether a row was observed from a completed value. Writing $Z_i^R = (Z_{iv})_{v\in\mathcal{A}_R}$ and $\rho_T(z) = P(R_i^T = 1 \mid Z_i^R = z)$, this mode requires

$$R_i^T \perp \mathcal{U}_i^T \mid Z_i^R, \qquad 0 < \epsilon_R \le \rho_T(Z_i^R) \le 1 \quad \text{almost surely}, \qquad T \in \{M, Y\}. \tag{16}$$

The weights are built from an estimate rather than from ρ_T itself, and that estimate is produced by a further cross-fitting pass inside each outer training fold. The fold I_{-k} is split into at least two non-empty inner folds, and the value $\widehat{\rho}_T^{(-k)}(Z_i^R)$ attached to a training row comes from an inner-training subset that excludes that row, so it is an assembled inner out-of-fold prediction and not a fit on all of I_{-k} read back on its own rows. Inner cross-fitting that cannot be executed for want of rows makes the weighting mode unavailable. With $\bar{R}_T^{(-k)}$ the observation share in the fold, the unnormalised stabilised weight is

$$\widetilde{w}_i^{T,(-k)} = R_i^T \, \frac{\bar{R}_T^{(-k)}}{\operatorname{clip}\{\widehat{\rho}_T^{(-k)}(Z_i^R), \epsilon_R, 1\}}, \tag{17}$$

normalised to mean one over the observed training rows. The fold reports the clipping fraction, the share of rows whose raw estimate fell below ϵ_R, and the effective sample size of the normalised weights, and the path-level record carries the smallest raw estimate, the largest clipping fraction and the smallest weight effective sample size across folds. A fold with no observed training row for a nuisance is assigned an effective sample size of zero and fails the weighting. Clipping is a numerical truncation and certifies nothing about positivity in the population.

The correction reaches the training law of the mediator and outcome models and stops there. It requires in addition that every member of C_π be raw-observed on every estimation row, since the outer average is taken against the raw law of C_π and no weight in this mode restores that law when the adjustment record is itself incomplete. This mode provides no outer weighting by an observation propensity for C_π, which is why raw missingness in the adjustment record moves the whole path to fold-wise completion rather than being repaired in place. A weighting failure of any kind has the same consequence, and it moves the path rather than a fold, so a per-fold fallback to a completed fit is not an equivalent procedure and is not permitted. The exposure propensity and the observation propensity are different objects with different arguments and different roles, and neither weights the final estimate.

C.7 A DISTINCTION THAT IS NOT A MATTER OF LABELLING

Giving completed rows positive weight inside the weighted objective is not the same procedure as fold-wise completion, and the two are never reported under one label. The first leaves the nuisance target defined against the raw law while training on a mixture whose composition is chosen by the data, and neither the missingness-at-random condition nor a correct observation model on its own controls what that mixture estimates. The second changes the nuisance target and the outer law openly, records the fixed fold-specific completions it is conditional on, and is reported with the completed-data label. A record therefore names the mode it ran under, and a completed-data result is a defined downgrade rather than a deletion of every number the path produced. That label is never upgraded afterwards, since recovering a full-data quantity from such a fit would take a different estimator with its own stated conditions.

C.8 THE THREE MISSING-DATA MODES SIDE BY SIDE

Each path uses exactly one of the three, and a path meeting none of them is recorded as not run.

Mode	Nuisance training	Outer target
Fully observed	raw rows, unit weights	raw empirical law of C_π
Observed-row weighting	raw-observed rows, observation-IPW	raw empirical law of C_π
Fold-wise completion	fixed per-fold completions	completed empirical law of C_π

D COUNTERFACTUAL INTEGRATION AND THE ESTIMATION SAMPLE

This appendix specifies how equation 5 is evaluated for a discrete and for a continuous mediator, which diagnostics accompany each route, what the fixed estimation sample enforces, and what the reported variance does and does not cover. Every quantity is defined on the estimation sample $\mathcal{I}_{\text{est}}$ with the cross-fitted nuisances of Appendix C.

D.1 DISCRETE MEDIATORS

For a mediator with finitely many categories the integral is a finite sum and is evaluated by exact enumeration. Writing $\widehat{p}_M^{(-k)}(r \mid x, c)$ for the probability mass induced by the training-fold mediator model over the fixed category coding, the row contribution is

$$\widehat{\gamma}_{ab,i} = \sum_{r=0}^{K_M-1} \widehat{p}_M^{(-k(i))}(r \mid x_b, C_{\pi,i})\, \widehat{\mu}_Y^{(-k(i))}(x_a, r, C_{\pi,i}), \tag{18}$$

which for a binary mediator is the two-term average over $r \in \{0, 1\}$. The coding of the category set is fixed together with the estimation sample, so the same integer labels the same category in the mediator model, the outcome model and the record.

Nested support is quantified rather than assumed. With a prespecified $\epsilon_M > 0$,

$$v_{M,ab}^{\text{disc}} = \frac{1}{n_{\text{est}}} \sum_{i \in \mathcal{I}_{\text{est}}} \sum_{r=0}^{K_M-1} \widehat{p}_M^{(-k(i))}(r \mid x_b, C_{\pi,i})\, \mathbf{1}\{\widehat{p}_M^{(-k(i))}(r \mid x_a, C_{\pi,i}) < \epsilon_M\} \tag{19}$$

measures how much of the mediator mass assigned under one exposure level sits on categories the other level barely reaches. A positive value raises a mediator-support warning and changes no number, since a finite-sample threshold on estimated probabilities does not demonstrate that population support inclusion fails. The estimate continues to be emitted whenever the required probabilities and outcome predictions are finite.

This enumeration covers a single discrete mediator. A path carrying several mediators, or mediators of mixed type, has no integration rule defined here and is recorded as out of scope and not run. That state says the present estimator does not cover the case, and it is never recorded as a failure of identification.

D.2 THE QUADRATURE BANK FOR A CONTINUOUS MEDIATOR

A continuous mediator cannot in general be handled by substituting its conditional mean, because $\mathbb{E}\{g(M) \mid X, C\}$ need not equal $g\{\mathbb{E}(M \mid X, C)\}$ when the mediator channel is not affine. The conditional-mean shortcut is exact for every mediator law only when the outcome regression is affine in m, which the fixed architecture of Appendix C is not, so the full integral is evaluated on every row. The mediator law is represented through the location–residual approximation $\widehat{F}_M^{(-k)}(m \mid x, c) = \widehat{F}_\varepsilon^{(-k)}\{m - \widehat{\nu}_M^{(-k)}(x, c)\}$, which maintains the working model $M = \nu_M(X, C_\pi) + \varepsilon$ with a residual law that does not vary with (X, C_π). That model is a working restriction rather than a general conditional-distribution learner, and a path relying on it carries the working-distribution label in its status fields. Where the restriction is not credible the remedy is a genuine conditional-distribution learner rather than a relabelling of this one.

The integration uses a registered quadrature size L, fixed at 64, and a registered node-construction rule. The training residuals of fold k are the observed mediator values of its training rows minus the fold model's fitted values on those rows, on the mediator's original scale. Write N_k for their number and L_k for the size of the bank, equal to L when $L < N_k$ and to N_k otherwise. Within each outer training fold one bank of nodes and weights $\{(\varepsilon_\ell^{(-k)}, \omega_\ell^{(-k)})\}_{\ell=1}^{L_k}$ is built from the residual empirical

law of that fold, with $\omega_\ell^{(-k)} \geq 0$ and $\sum_\ell \omega_\ell^{(-k)} = 1$. When $L < N_k$ the nodes are mid-quantiles of that law. Writing $F_\varepsilon^{(-k)}$ for the weighted empirical distribution of the training-fold residuals,

$$\varepsilon_\ell^{(-k)} = \inf\{\varepsilon : F_\varepsilon^{(-k)}(\varepsilon) \geq (\ell - \tfrac{1}{2})/L\}, \qquad \omega_\ell^{(-k)} = 1/L, \qquad \ell = 1, \ldots, L,$$

so the nodes carry equal weight and the residual law enters through where the positions fall. Taking those positions on the weighted distribution is what carries the weights of observed-row weighting into the bank, since a down-weighted residual advances less cumulative mass and is selected at fewer positions. When $L \geq N_k$, or L is left unset, no reduction is performed and the bank is the residual empirical measure itself with its own weights, so $L_k = N_k$. The bank is built from the training fold alone and is not indexed by validation row, so it cannot become mechanically coupled to $C_{\pi,i}$, and it is recorded with its nodes and weights and the mode that produced them. Each validation row is then evaluated over the whole bank,

$$\widetilde{M}_{b,i,\ell} = \widehat{\nu}_M^{(-k(i))}(x_b, C_{\pi,i}) + \varepsilon_\ell^{(-k(i))}, \qquad \widehat{\gamma}_{ab,i} = \sum_{\ell=1}^{L_{k(i)}} \omega_\ell^{(-k(i))}\, \widehat{\mu}_Y^{(-k(i))}\big(x_a, \widetilde{M}_{b,i,\ell}, C_{\pi,i}\big). \quad (20)$$

One bank serves the reference and comparison arms rather than one being drawn for each, which removes a source of noise that would otherwise enter the contrast without carrying information about it. A finite bank still carries a numerical approximation error when the outcome model is nonlinear in m, sharing it between the arms does not guarantee that this error cancels in the contrast, and the bias identity of Appendix C assumes exact integration.

The residual law behind the bank is the one the active missing-data mode delivers. In the first mode the bank comes from the raw residual law of the training fold. In the second it comes from weighted quantiles of the observation-weighted residual law, using the same weights that trained the mediator model. In the third it comes from the residual law of the completed training fold, and the completed-data label carries over to every quantity built on it, the row contributions and the contrasts included.

D.3 Mediator support and extrapolation

Generated mediator nodes may fall outside the support the outcome model was trained on, and that is measured under the quadrature weights rather than counted by row. Let $\widehat{\mathcal{S}}_{M,a}^{(-k)}(c)$ be the arm-a mediator support envelope built on the training fold by the registered rule. Without a registered rule the mediator support is recorded as not assessed and the fraction below is not computed. The extrapolation fraction is

$$v_{M,ab}^{\mathrm{cont}} = \frac{1}{n_{\mathrm{est}}} \sum_{i \in \mathcal{I}_{\mathrm{est}}} \sum_{\ell=1}^{L_{k(i)}} \omega_\ell^{(-k(i))}\, \mathbf{1}\big\{\widetilde{M}_{b,i,\ell} \notin \widehat{\mathcal{S}}_{M,a}^{(-k(i))}(C_{\pi,i})\big\}. \quad (21)$$

A positive value raises a support and extrapolation warning, and the effect stage retains the numerical result whenever the outcome learner still returns finite predictions there. The boundary is drawn at computability. An empty residual bank, undefined weights, an integral that cannot be evaluated or a non-finite prediction makes the quantity not computable, and that is the only route by which a support problem makes the effect stage drop a number. Whether the reporting policy releases a retained number depends on the support rule (Appendix E).

D.4 The estimation sample

The estimation sample is fixed by the split plan before estimation, and its rows are recorded. Define the finite-row set $\mathcal{I}_\pi^{\mathrm{fin}} = \{i \in \mathcal{I}_{\mathrm{est}} : \widehat{\gamma}_{ab,i} \text{ is finite for all } (a,b) \in \mathcal{J}_{\mathrm{main}}\}$, over the three nested means the decomposition uses. The requirement covers those three and no others, so $\widehat{\eta}_{01}$ is computed and reported as a diagnostic of the alternative decomposition and never blocks an output it does not enter. A path-level mean is released only when $\mathcal{I}_\pi^{\mathrm{fin}} = \mathcal{I}_{\mathrm{est}}$ and the row identifiers match the recorded sample. Otherwise no path-level mean is released for that population, and the reason records which condition failed.

Two rules follow and both are enforced rather than recommended. Support and extrapolation warnings never delete an estimate that is finite and uniquely defined. Rows are never dropped after the estimation sample is fixed, because removing the non-finite rows and averaging over what remains reports a different population under the name of the fixed one. A finite-row descriptive summary

may still be useful and may be reported, provided it is labelled with its own row-population identifier so that it cannot be read as the fixed estimate.

D.5 THE ROW-SCORE VARIANCE DIAGNOSTIC

The row-level contrast contributions are $\psi_{i,\pi}^{\mathrm{NDE}} = \widehat{\gamma}_{10,i} - \widehat{\gamma}_{00,i}$, $\psi_{i,\pi}^{\mathrm{NIE}} = \widehat{\gamma}_{11,i} - \widehat{\gamma}_{10,i}$ and $\psi_{i,\pi}^{\mathrm{TE}} = \widehat{\gamma}_{11,i} - \widehat{\gamma}_{00,i}$, and each plug-in contrast of Section 3.6 is their mean over the estimation sample $\widehat{\theta}_{\pi,\kappa}^{\mathrm{plug}}$. The reported variance is the sample variance of those contributions divided by n_{est},

$$\widehat{V}_{\pi,\kappa}^{\mathrm{row}} = \frac{1}{n_{\mathrm{est}}(n_{\mathrm{est}}-1)} \sum_{i \in \mathcal{I}_{\mathrm{est}}} \left(\psi_{i,\pi}^{\kappa} - \widehat{\theta}_{\pi,\kappa}^{\mathrm{plug}}\right)^2, \qquad \kappa \in \{\mathrm{NDE}, \mathrm{NIE}, \mathrm{TE}\}, \tag{22}$$

which requires $n_{\mathrm{est}} \geq 2$. The total-effect entry is computed from $\psi_{i,\pi}^{\mathrm{TE}}$ directly and never as the sum of the direct and indirect variances, a substitution that would discard the covariance between them.

The quantity is conditional on the fixed set, the fixed architecture, the fold assignment and the completion draws, and it propagates none of them. It also propagates neither the estimation of the nuisances nor of the observation propensity, neither the numerical integration, the structural checks made before outer-fold fitting, nor the selection of the path and its adjustment set. CHARM reports a cross-fitted plug-in for the natural effects rather than an efficient-influence-function estimator, so equation 22 is read as a dispersion diagnostic and supports no coverage statement.

E THE PATH-LEVEL RECORD

This appendix gives the record behind Section 3.7, the admissible values of its nine status fields, the two denominators and the rule that separates them, and the integrity grades. The design principle throughout is that a number is removed only when it cannot be computed or cannot be uniquely attributed, and that every other concern moves a label instead.

The record is specified in full, but the runs of Section 4 recorded only part of it: the fixed set and its identifiers, the missing-data mode, the exposure and mediator support rules, the integration details, the estimation sample and the effect-status labels of Appendix C. The remaining fields, the external pre-exposure check and the integrity grades are specified here and were not recorded in those runs. Release is handled separately from computation: an estimate is marked as released once its exposure and mediator support have been assessed and found eligible. No mediator support rule was registered for the end-to-end campaign, so this release rule does not apply there, and Table 3 reports the effect-stage estimates directly. The release rule reads no estimate and changes no number.

E.1 THREE KINDS OF RECORD

A path whose estimator did not run to completion produces a run record rather than an effect record. It carries the path, the adjustment source, the computability state, a reason, the fixed set where one exists, and the family status. Missing design input, a compiler that did not execute, a bridge that issued no set, an exhausted fit budget, cross-fitting without a non-empty training fold, a solver that failed and an estimator not implemented for the case all land here under their own reasons, and the run continues with the remaining paths. A bridge that issues no set records one of its reasons, no registered path clearing the stability threshold, an obstruction in a supporting graph, fewer than three supporting graphs in a panel, a proposal that fails certification on a supporting graph, or a tie that only labels would break. Keeping this record separate is what stops a path that was never attempted from later reading as one whose estimate was withdrawn.

A path that fixed its adjustment set and entered estimation produces the effect record. The computability state decides between the two: not run goes to the run record and numeric or undefined to the effect record, so an estimator that entered estimation and then failed is recorded in the run record under its own reason. It holds the three plug-in contrasts and their row-score variance diagnostic, the reported path stability score with its denominator, the fixed C_π and its adjustment source, the external pre-exposure check, the panel certification and obstruction rates, the support vector, the nine status fields, the denominator status and the minimum information needed to rebuild the analysis. The support vector collects the exposure violation fraction, the per-arm effective sample sizes, the mediator support diagnostics of Appendix D, and, under observed-row weighting, the smallest raw

observation propensity, the clipping fractions and the observation-weight effective sample sizes. A component that does not apply to the active mode is recorded as not applicable and never as zero.

The existence configuration of Appendix B runs no estimator by design, so it produces a third record holding the candidate family, the per-graph candidate sets and the existence, common-set and obstruction rates by panel. Its effect scope is diagnostic only, and it fabricates neither an effect vector nor a missing value standing in for one.

Fields that have not been generated are recorded as such. A missing value appears only inside an effect record whose computability state is undefined, marking the numerical fields the estimator was unable to produce. The distinction matters because the two look identical in a table and mean different things about what was attempted.

E.2 The nine status fields

The record carries nine status fields, and they vary independently rather than forming a ladder.

Field	Admissible values
Computability	numeric, not run, undefined
Adjustment source	design fixed, compiled, existence audit
Causal interpretation	assumption conditional, trusted-graph conditional, withheld, not assessed
Family compatibility	full, partial, none, diagnostic incomplete, not assessed
Missing-data mode	fully observed, observed-row weighting, fold-wise completion
Support	supported by diagnostics, support sensitive, model extrapolation, not assessed
Working law	registered, working distribution, location-shift conditional, not applicable
Scope	in scope, out of scope
Integrity	verified, degraded, undefined

Computability is numeric when the target, the row population, the fixed set, the integral and the contrasts are all defined and finite. It is undefined only when roles, intervention levels and estimand conflict, when the estimation sample is empty, when a required integral or normalising denominator is undefined, or when the row population and adjustment set behind a number cannot be determined. Every other failure to produce a number is not run, and none of them is ever recorded as a failure of identification. A budget, a solver, a missing design-supplied set, a compiler that did not execute, a bridge that issued no set, cross-fitting without a usable training fold and an unimplemented multi-mediator routine are all in that second class.

E.3 Causal interpretation

The causal status is decided by the evidence an analyst maintains and never by a graph rate. The external pre-exposure check is a three-valued field recorded alongside the record, and it is not an adjustment branch, since all three of its values attach to the same fixed C_π and the same numerical functional. A record is labelled an assumption-conditional natural effect when the analyst explicitly maintains (A1) to (A5) for the realised path and set together with the missingness and nuisance conditions in force, which is a conditional claim of the same kind any identification theorem makes and is not something the compiler establishes. It is labelled trusted-graph conditional when an external graph is declared, the fixed set is causally admissible on it, and the same conditions are maintained. A failed check, or a trusted graph that rejects the route, leaves the same computable functional in place under a withheld label, and removing variables afterwards while keeping the original identifiers is not an available response. Where nothing is attested, no trusted graph is declared and the assumptions are not explicitly maintained, the field reads not assessed. Saying elsewhere that the causal label is withheld by default means that no causal label is issued; in this field that default reads not assessed, and the value withheld is reserved for a failed check or a trusted graph that rejects the route. By default the label is withheld, no internal state of the set raises it, and no certification rate on the sampled family, however high on either panel, substitutes for this evidence.

E.4 Sampled-family compatibility

The family field reads full when both panel certification rates equal one, partial when at least one is positive without both reaching one, and none when both are zero. It reads diagnostic incomplete

when either panel supports the path nowhere or the denominator is operational, and not assessed when no set was issued. These labels describe the supplied family through the panel diagnostics of Appendix B and are not held-out guarantees, so a trusted-graph-conditional record with no family support is coherent, and so is a not-assessed causal label with full family support. Candidate existence and candidate ambiguity are reported as separate diagnostics and do not map onto this field.

E.5 The two denominators

The path stability score is $\widehat{\mathrm{PSS}}(\pi) = m_\pi/S$ over the registered ensemble, with m_π the number of draws containing both path edges. When the list of supporting graphs, the graph identifiers, the stored records or the panel assignment cannot be recovered, the reported triple switches to the operational one built from the effective number of graphs, and the denominator status records that substitution. In that state the registered panel rates and the obstruction rates are recorded as not assessed and the family field reads diagnostic incomplete. Operational counts and rates may still be reported for the graphs that survive, and they are never renamed as the complete family diagnostic. Restoring a complete-family label by substituting one denominator for the other is prohibited, as is deleting the missing indices and presenting the resulting proportion as the registered one. A numerical effect record survives an incomplete graph record whenever the set actually used, the effect inputs and the estimand remain uniquely reconstructible, and what degrades is the family and integrity labels rather than the number.

E.6 Why the panel rates carry no interval

A binomial Monte Carlo interval on a certification rate would describe the sampling fluctuation of a fixed candidate evaluated on draws it never influenced. The compiled set does not meet that condition, because both panels contributed node evidence to the compilation, so the rates are reported as raw counts and rates with their reason-code histograms and nothing more. An interval would become available under a different design in which the set is fixed before a panel is drawn and the graphs of that panel are drawn independently under the fixed sampler, and adopting it would mean giving up the full-ensemble evidence that the recall-first policy of Appendix B depends on. The same construction is unavailable for any rate defined as a maximum over candidates, since the maximum is not the frequency of a fixed candidate and the binomial form does not apply to it whatever the panel design.

E.7 Integrity and traceability

Integrity is verified when the bindings agree. It is degraded when a stored record is incomplete while the data, the effective adjustment set, the row population and the estimand still reconstruct uniquely, in which case the number stands and the claim is capped as exploratory. It becomes undefined only when a conflict leaves the estimand, the selected set or the computation inputs no longer uniquely determined. Recomputing a policy, substituting a set, deleting a graph or exchanging the row population in order to recover a higher label is prohibited, which is what makes the grades informative.

Each path stores what is needed to rebuild it: the path and its intervention levels, the estimand and effect method, the adjustment set with its source and the graph evidence behind a compiled one, the state of the external pre-exposure check, the path score and threshold, the data, the fold-specific completions, the estimation sample, the split and the architecture, the ensemble sizes with the supporting graphs and panel assignment, and the missing-data, support, working-law and scope labels. Checks run on these stored records rather than on the source code, so a claim is checked against what a run recorded rather than against what an implementation appears to do.

F Experimental protocols

This appendix collects the protocols, the metric definitions and the data-generating processes behind Tables 1, 2 and 3. Everything a table reports is defined here once, and the individual experiment sections state only what their result turns on.

F.1 CONVENTIONS SHARED BY THE EXPERIMENTS

The truth is fixed and hidden before a run starts. The generating graph, the true path and the true effects are written once, and no stage of the pipeline reads any of them, the discovery screen of Section 3.2 included, except where an experiment supplies them by design, as the mechanism panel does when it hands the true path and a valid adjustment set to every method. Errors are always taken within a replicate before being averaged, so a positive error on one replicate cannot cancel a negative one on another, and the spread reported beside a mean describes replicate-to-replicate variation rather than an interval, with each table stating whether it is a standard deviation or the standard error of the mean. Denominators are the planned ones. A replicate that produces a fallback is counted in the denominator that was registered rather than removed from it, so an expected fallback lowers a rate instead of disappearing from the record.

Every bridge hyperparameter is registered before any effect estimate is examined, at $S{=}30$ graphs, a diagnostic pool of $Q{=}12$ and a set-size cap of $K{=}2$. The two panels of Appendix B are the fixed assignment A = sampled graphs 1–15 and B = graphs 16–30 under the same registration.

F.2 STRUCTURE DISCOVERY

The discovery experiment evaluates the stage on synthetic DAGs of increasing width at fixed sample size, with $p \in \{500, 1000, 2000, 5000\}$ at $n{=}10^4$, and $R{=}10$ replicates at p up to 2000 and $R{=}3$ at $p{=}5000$. The data-generating process is given below. For each replicate every method receives the same completed table, built once by the K-MICE completion of Section 3.2 under the settings of Appendix G.2, so that differences between rows reflect the search rather than the completion, and the whole sample of each replicate serves discovery. Every method is scored against the same true graph by the same code. CHARM runs the stage of Section 3.2 under the settings of Appendix G.2, and the comparison methods run as described below. Each accuracy entry of Table 1 is the mean over replicates of a quantity computed within one replicate, and the value after $\pm$, where one is shown, is the standard error of that mean.

Discovery data-generating process. Each replicate draws a DAG over p variables from an ordered Erdős–Rényi model with expected in-degree $k{=}3$, so that each forward pair carries an edge with probability $2k/(p-1)$, and embeds two parallel mediation paths from the exposure to the outcome, each with at most two chained mediators. Edges added by the embedding or by a connectivity repair are counted apart from the base graph and belong to the true graph, and no graph is rejected or redrawn for its density. Each node is continuous, binary or categorical with at most five levels, with probabilities 0.6, 0.3 and 0.1. Every node follows an additive-noise structural equation in its parents with linear coefficients of magnitude in $[0.2, 1.0]$ and random sign. A continuous node is nonlinear with probability 0.5, adding a quadratic, sine or $\exp(-|\cdot|)$ term in one parent with coefficient magnitude in $[0.3, 1.5]$, and occasional pairwise product terms appear. Binary and categorical nodes are drawn through logistic and category-specific multinomial logit links. The noise is Gaussian, exponential or Student-t with probabilities 0.6, 0.2 and 0.2, with standard deviation in $[0.5, 1.5]$ and a target signal-to-noise ratio of 4. The raw direct effect is scaled into $[0.35, 0.50]$ and the raw indirect effect into a band of $\pm 10\%$ around a midpoint calibrated for each replicate towards a realised $|\mathrm{NIE}|$ of 0.25, by at most a factor of 6. The true effects are computed from $20{,}000$ Monte Carlo draws, and a candidate is admitted only if its realised $|\mathrm{NIE}|$ lies in $[0.15, 0.35]$ and $|\mathrm{TE}| \geq 0.10$; the realised direct effect is recorded but not gated. Missingness is MAR at an overall rate of 10%, spread over 90% of the columns. For each such column, five conditioning columns are drawn without replacement from the columns that are never made missing, whatever their role. Their standardised values enter a logistic score with standard normal coefficients, rescaled to unit standard deviation, and the intercept is solved by bisection to meet the target rate. Table 4 lists the design and the seeds.

Two scope statements govern how this panel is read. No function of the true graph reaches any method, while the registered roles of X, M and Y reach all of them. Every method receives the role tier as forbidden edges, and the screen of CHARM also uses the roles for its role additions. And accuracy is scored against the true graph over all $\binom{p}{2}$ unordered pairs, so a true edge that the candidate screen never admitted is counted as a false negative rather than excluded from the denominator. That convention is what puts a screened search and a full search on one scale.

Table 4: Design of the discovery benchmark (n=10^4 at every scale). Each replicate is identified by its structural seed s, which seeds the graph; the mediation embedding, the node types, the structural equations, the complete rows, the missingness mask and the Monte Carlo ground truth use s+1 to s+6. The seeds are those of the admitted candidates. The registered design has three replicates at p=5000, which its cost fixed in advance.

p	R	Edge prob.	Expected base edges	Structural seeds
500	10	0.0120	1500	7001, 7002, 7007, 7008, 7009, 7010, 7017, 7025, 7029, 7036
1000	10	0.0060	3000	8005, 8006, 8008, 8010, 8011, 8019, 8022, 8023, 8030, 8031
2000	10	0.0030	6000	9002, 9003, 9006, 9007, 9010, 9013, 9014, 9016, 9021, 9026
5000	3	0.0012	15000	10001, 10002, 10004

F.3 Discovery metric definitions

Adjacency F1 is the harmonic mean of precision and recall on the undirected skeleton, scored over every one of the $\binom{p}{2}$ pair positions. A true edge that the candidate screen never admitted is a false negative rather than an excluded case, which is what puts a screened search and a full search on one scale.

Normalised structural Hamming distance counts, over all pairs, the predicted adjacencies absent from the truth, the true adjacencies not predicted, the true adjacencies given the reversed direction and the true adjacencies left without a committed direction, and divides the total by the number of true edges. It is the one column that charges all four kinds of error in a single number, and it separates methods that adjacency F1 ties, as it does for PC-Stable and BOSS at p=500. A value above one means more structural errors than the graph has edges, which holds for PC-Stable at every scale of Table 1.

Orientation F1 is computed over ordered pairs against the directed truth. A detected adjacency without a committed direction is entered in both directions instead of guessed, so on a true edge it adds one correct and one incorrect direction, and the column is read together with the next two. The abstention rate is the share of detected true edges left without a committed direction, and the committed accuracy is the share of committed directions on true edges that match the truth, against a chance level of one half. A low abstention rate beside a committed accuracy near one half describes a method that guesses, which is the pattern of PC-Stable in Table 1.

Pooled path F1 is a directed F1 on the role subgraph, the submatrix over X, the candidate mediators and Y. Within one replicate it pools true positives, false positives and false negatives over all ordered role pairs before the F1 is taken, with an uncommitted pair entered in both directions as above, and the table reports the mean of the per-replicate values with the standard error of that mean. The role subgraph holds few edges, so one edge moves the value of a replicate by a large step, and the standard errors in this column are between about six and twenty times those of adjacency F1 in the same cell.

Pairs searched is the number of unordered pairs in the searchable domain handed to the search, as a share of $\binom{p}{2}$. A search without a screen covers every pair by construction, and a smaller share records a property of the design rather than a gain in accuracy.

Wall time is the median over replicates of the search process, from handing the completed table to the searcher to the scored output, and the screen of CHARM is built in a separate step whose time is not included. Recorded memory is a reading for the search process, given as the median over replicates or, where no median was recorded, as the midpoint of the smallest and largest reading. At p=5000 it is the peak live heap the Java collector logged during the search, and at the smaller scales it is the resident set the operating system reported when the search returned. Both readings are lower bounds on the memory the search held at its peak.

F.4 Comparison methods for discovery

The three comparison methods receive the same completed table, the same role tier as forbidden edges and the same scoring code as CHARM, and each searches every pair at the configuration listed in Table 5. FGES-Full runs the greedy equivalence search and score of CHARM over every pair,

at its own penalty discount and degree bound. PC-Stable runs the PC procedure with its order-independent adjacency phase (Colombo & Maathuis, 2014) and a degenerate-Gaussian likelihood-ratio test. BOSS runs a best-order score search (Andrews et al., 2023) under the degenerate-Gaussian score with a single start and without its backward equivalence phase, the defaults of its Tetrad implementation (Ramsey et al., 2018). The output of every method is read through the endpoint convention of Appendix G.2, so an adjacency whose direction a method declined to commit is treated identically across the four rows. The configuration of CHARM is set per scale, with a penalty discount of 4 on one thread up to $p{=}2000$ and a penalty discount of 8 on eight threads at $p{=}5000$, while PC-Stable and BOSS keep one configuration at every scale, and FGES-Full uses the configuration of CHARM up to $p{=}2000$ (Table 5). At that configuration FGES-Full exhausted a 196 GiB heap on all ten replicates at $p{=}2000$, after a median of 6.2 h (4.0 to 71.4 h) within a 72 h limit, and at $p{=}5000$ it exhausted the same heap on all three replicates at a penalty discount of 64 and a degree bound of 5. Every configuration was chosen by a manual search for each method on pilot replicates disjoint from those reported, weighing accuracy against whether the search completes within its time and memory budget, and each row of Table 5 is the best setting that search found for that method.

Table 5: Search configuration of each discovery method in Table 1. DG is the degenerate-Gaussian form, BIC the Bayesian information criterion and LRT the likelihood-ratio test. A dash marks a setting the procedure does not read. The last row, marked e2e, is the end-to-end setting of Section 4.3, which also adds every baseline-to-X, $\mathcal{M}$ or Y pair to the screened domain, orders all variables in five public tiers and builds the graph family of Section 3.3.

Method	Procedure	Score or test	Screen	Penalty discount	Max. degree	α / depth
CHARM	Greedy equivalence	DG BIC	yes	4; 8 at $p{=}5000$	10	–
FGES-Full	Greedy equivalence	DG BIC	no	4; 64 at $p{=}5000$	10; 5 at $p{=}5000$	–
PC-Stable	PC, stable adjacency	DG LRT	no	–	–	0.05 / 2
BOSS	Best-order score	DG BIC	no	8	–	–
CHARM (e2e)	Greedy equivalence	DG BIC	yes	4	32	–

F.5 THE MECHANISM PANEL

Mechanism panel: methods. Table 2 compares outcome-model choices under one estimation protocol, with the true path, a valid adjustment set and the same data supplied to every method. The methods do not share a fold assignment, and each follows its own fitting protocol. CHARM runs the cross-fitting scheme of Appendix C, the g-computation baselines cross-fit their nuisances over five folds of their own, with CV-selected choosing its outcome family in a separate three-fold stage, and PL-DML residualises on the adjustment set by cross-fitting. The *Oracle-form diagnostic* receives the family-specific correct functional form, is fitted once on the full sample without cross-fitting, and serves as a finite-sample reference diagnostic, not a strict error bound. CHARM fits the base class of equation 4, linear in the exposure and flexible in the mediator and in the adjustment set, without the interaction diagnostic, since the panel contains no exposure–mediator interaction. `CV-selected` chooses its outcome family by validation and is an adaptive comparator rather than a CHARM variant. Spline and MLP fix the outcome model to a spline basis and to one unrestricted network over (x, m, c) respectively; the latter lacks the additive factorisation of CHARM and also differs from it in its training settings. Parametric g-computation (Robins, 1986) and the partially linear `PL-DML` estimator (Robinson, 1988; Chernozhukov et al., 2018) are the two restrictive references. Two mediation-specific estimators were added after the panel had been run, on the same effect halves, adjustment sets, intervention levels and truths. Under binary exposure, mediation DML (Farbmacher et al., 2022) evaluates the efficient score of `causalweight` 1.1.5 with random-forest nuisances fixed before any effect error was inspected, five cross-fitting folds of its own and propensities clipped to $[0.01, 0.99]$, which touched at most 2.4% of rows; its direct and indirect effects are the $\theta(0)$ and $\delta(1)$ that match the NDE and NIE of Section 2. That estimator requires a binary exposure, so under continuous exposure we use generalised-propensity-score weighting (Huber et al., 2020), a weighting rather than a debiased estimator, with the package bandwidth rule.

Mechanism panel: data-generating process. The seven families share one structural core and differ only in the mediator–outcome channel. With standard normal C_1, C_2 and continuous ex-

posure $X = 0.25 + 0.3C_1 - 0.2C_2 + 0.75\varepsilon_X$, replaced under the binary profile by $X \sim \text{Bernoulli}[\text{expit}(\beta_0 + 0.3C_1 - 0.2C_2)]$ with $\beta_0 = -1.130$, the mediator is linear in every family,

$$M = 0.5X - 0.2C_1 + 0.2C_2 + 0.85\varepsilon_M,$$
$$Y = \varsigma\{0.20X + c_{\text{lin}}M + \gamma_k\phi_k(M)\} + 0.15C_1 - 0.15C_2 + \sigma_{Y,k}\varepsilon_Y,$$

with C_1, C_2 and the exposure innovation standard normal, the mediator and outcome innovations generalised Gaussian with shape parameter 4 and unit variance, and $\varsigma \in \{-1, +1\}$ a hidden effect sign. The six nonlinear families are fixed at η=0.95, the share of the indirect contrast carried by ϕ_k, which fixes c_{lin}=0.02 and the amplitudes below; the linear family is the η=0 member of the same path, with γ_k=0 and c_{lin}=0.40. Each $\sigma_{Y,k}$ is set separately for every family, exposure profile and effect sign so that SD(Y)=1, and every contrast is taken between the intervention levels x_0=0 and x_1=1. Each of the ten replicates is generated from its own random key, fixed before the run. Writing μ_M=0.25 and σ_M=$\sqrt{0.865}$ for the fixed centre and scale of M (denoted C and S in the released registry):

Table 2 name	Registry family	$\phi_k(m)$	γ_k
Linear	`linear_eta0`	— (carrier only)	0
Sin	`sin`	$\sin(m/0.6)$	0.9574
Tanh	`tanh_narrow`	$\tanh\{(m-\mu_M)/(0.20\sigma_M)\}$	0.5088
Ramp	`clipped_ramp`	$\text{clip}\{m-(\mu_M-0.5\sigma_M), 0, 0.75\sigma_M\}$	1.4636
RBF	`rbf_offcentre`	$\exp\{-(m-\mu_M-\sigma_M)^2/(2(0.8\sigma_M)^2)\}$	1.3946
Softplus	`softplus_right05_tau02_eta95_gg4`	$\tau_s \log\{1+\exp((m-a)/\tau_s)\}$	1.1455
Poly2	`raw_poly2_eta95_gg4`	m^2	0.7614

In the Softplus row a=μ_M+0.5σ_M and τ_s=0.2σ_M. **The panel contains no exposure–mediator interaction**: X enters the outcome only through 0.20ςX and ϕ_k is a function of M alone, so the true outcome mean lies inside the closure of the working class of equation 4 for every family. What the panel varies is the shape of the mediator channel g (Figure 3), not whether the additive separation holds; effect modification, which equation 4 cannot represent, is not tested here.

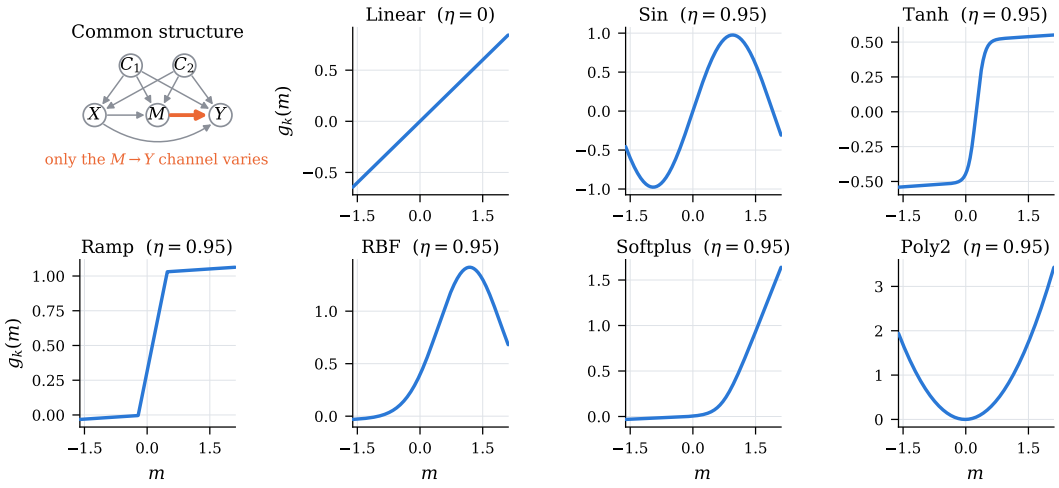


Figure 3: The seven mechanism families of Table 2. They share the causal structure on the left, the exposure and mediator equations, the direct effect and an indirect effect of 0.2, and differ only in the mediator–outcome channel $g_k(m) = c_{\text{lin}}m + \gamma_k\phi_k(m)$ of the table above, shown for a positive effect sign over the common reference window $\mu_M \pm 2\sigma_M$, which holds about 97% of the equal mixture of the two interventional mediator laws. In each nonlinear family ϕ_k carries 95% of the indirect effect. Vertical scales differ across panels. The curves describe the generating rule, not a fitted model.

Mechanism panel: metrics. Per-family cells give MAE = mean$_r|\widehat{\theta} - \theta^\star|$ over the R=10 replicates, taken within replicate so that signed errors cannot cancel, and reported separately for the

indirect and direct contrasts. *Macro* is the equal-family mean of those values. L_{rm} is the relative mean error of equation 6, reported as a percentage, *Mean* is its equal-family mean and *Worst* its largest family value with the attaining family named.

F.6 END-TO-END COMPOSITION

The end-to-end experiment runs discovery and estimation in one pass, since neither of the first two panels establishes that the two stages compose. Every replicate starts from a p-dimensional matrix, runs discovery under the end-to-end settings of Appendix G.3, selects the top-ranked learned path, compiles one adjustment set with no user-specified set, and hands that set to the effect stage of Appendix C. The reported campaign is $(p, n_{\mathrm{total}}, R) = (500, 10^4, 40)$, fresh and complete-data, with disjoint discovery and effect halves and an effect half of 5000 rows. Its registered path carries the sine mediator–outcome channel of the mechanism panel at η=0.95 under continuous exposure. Its seeds were drawn before generation and are disjoint from a registered list of earlier seeds, the plan and code were fixed before any data existed, and the hidden truth was revealed only after every replicate had finished and its outputs had been stored. The four registered strata, ordinary, near-instrument, collider-descendant and strong exposure-descendant, contribute ten replicates each, and each registered path carries two designed adjustment variables (Figure 4). The structural seeds are 260906101–260906140, assigned to the strata in this order in rotation, and each replicate also uses its own random key, fixed before the run. The effect stage runs the estimation protocol of Appendix C with five cross-fitting folds, an outcome network trained for sixty epochs at batch size 256 in place of the fifty epochs and batch size 128 listed there, a linear conditional mean for the continuous mediator, and a sixty-four-node quadrature bank. No mediator support rule was registered for this experiment, so mediator support is recorded as not assessed (Appendix D). Each row of Table 3a is a separate campaign, and campaigns are not pooled. Two further campaigns run the same design at larger width, at $(p, n_{\mathrm{total}}, R) = (5000, 10^4, 5)$ and $(10^4, 2 \times 10^4, 5)$, with structural seeds 260906501–260906505 and 260906701–260906705. Neither is stratified, so every replicate there is ordinary, and a replicate takes a median of 9.25 and 63.09 hours from start to exit.

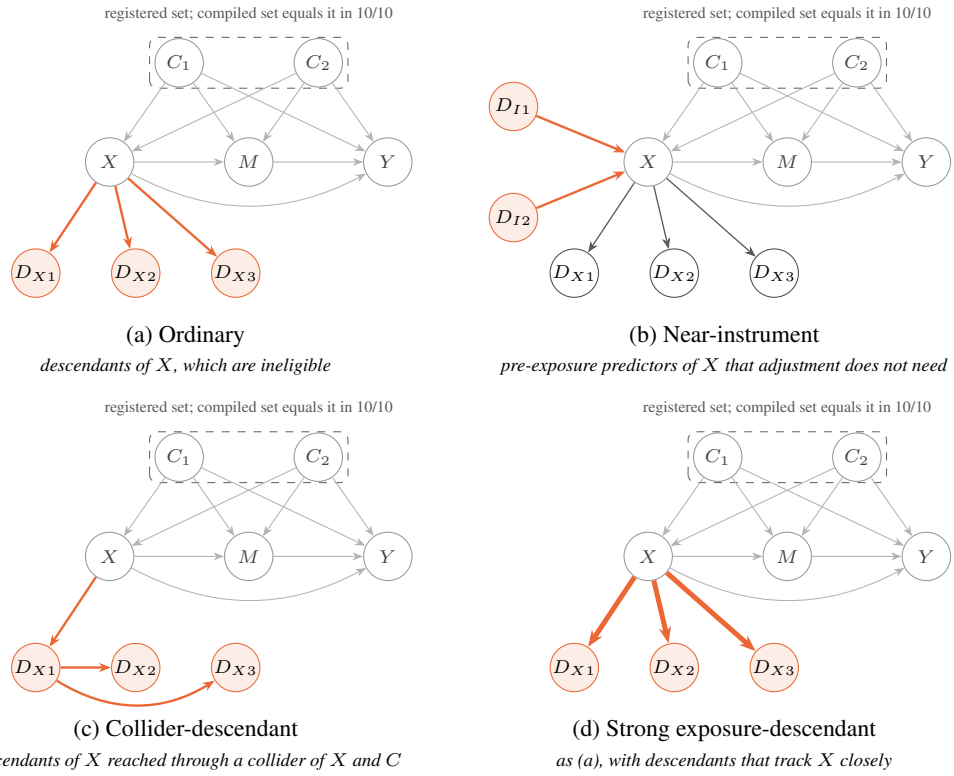
The reported denominator is the planned one, so a replicate that received no set or whose path selection missed the target is counted rather than dropped, and the target-path and graph-admissibility columns are shares of the registered R. The compiled and supplied sets coincide in every replicate, so the matched supplied-set replay of Appendix I, which reruns a replicate on the same rows, path and folds with the design-supplied set, returns identical estimates. MaxAE is the maximum over replicates and estimands within a row rather than a mean.

G THE SCREENED SCORE-BASED DISCOVERY STAGE

This appendix first specifies the MICE completion backend and our memory-reduced K-MICE implementation, then records the screened discovery configuration used in Table 1, whose inputs that K-MICE completion produces. The availability of a completion backend does not establish that it was enabled in a reported run.

G.1 MICE AND THE MEMORY-REDUCED K-MICE IMPLEMENTATION

The inherited method and our changes. Chained-equation imputation updates one incomplete variable at a time from conditional models fitted to the other variables (van Buuren & Groothuis-Oudshoorn, 2011). Our backend uses `IterativeImputer` with Bayesian ridge regression. Its option to sample a bounded set of predictors by correlation is an existing scikit-learn feature (scikit-learn developers, 2024); we call this configuration K-MICE to distinguish it from using all other columns. The name denotes a predictor budget, not a new family of conditional distributions or a nearest-neighbor method over rows. Our changes concern execution: a streaming implementation avoids retaining the full correlation matrix, and an optional one-shot policy releases conditional models after their current prediction. Neither changes the intended conditional-update or neighbor-sampling rule. Algorithm 2 specifies the completion step, with the optional changes marked explicitly.



(a) Ordinary
descendants of X, which are ineligible

(b) Near-instrument
pre-exposure predictors of X that adjustment does not need

(c) Collider-descendant
descendants of X reached through a collider of X and C

(d) Strong exposure-descendant
as (a), with descendants that track X closely

Figure 4: Generating structure of the four registered strata of the p=500 end-to-end campaign. Each stratum surrounds the same target path with variables, shown in colour, that a selector reading associations could pick by mistake, and the compiled set equals the registered set $\{C_1, C_2\}$ (dashed box) in all ten replicates of every stratum (Tables 9 and 10). The grey core, $C_1, C_2 \to X, M, Y$ with $X \to M \to Y$ and $X \to Y$, is common to all strata. Each D_X variable also has C_1 and C_2 as parents, except that the near-instrument stratum drops $C_2 \to D_{X2}$ and $C_1 \to D_{X3}$; the strong exposure-descendant stratum keeps the ordinary edges with a stronger dependence of each D_X on X. No stress variable is a parent of M or Y. Not drawn: the background graph, a second target path with a linear mediator channel, and seven variables that are children of C_1 or C_2 or have no parents. The two larger campaigns use the ordinary structure.

Fit scope, encoding and initialization. This subsection and Algorithm 2 specify the discovery Step 1 wrapper. The effect-stage adjustment imputer shares its chained-update mechanism, but not its preprocessing or one-shot execution scope; Appendix C specifies that separate procedure. Here n and p denote the rows and columns supplied to one imputer. Let $R_{ij} = 1$ indicate an originally observed entry. The caller supplies the discovery rows; the imputer does not choose the discovery/estimation partition itself. Continuous columns are standardized using their observed fit-scope entries. Binary and categorical variables retain their registered integer codes. No complete ground-truth table is used to fit these transformations, and row identities and order are preserved. A column with no observed fit-scope entries is outside this completion procedure.

Missing entries are initially replaced by their observed column means, giving $\widetilde{Z}^{(0)}$. This is an initialization for iterative fitting and is overwritten by the sweeps. The column order is fixed by increasing training missingness. Each sweep uses the most recently updated values of the other columns. The default iteration cap is $T_{\text{imp}} = 10$; with posterior sampling enabled, the implementation executes the configured sweeps rather than applying the deterministic imputer's convergence early-stop rule. For a single-column input there are no predictors: the implementation skips conditional fitting and returns the initialized values after the same output projection and checks.

Conditional fitting and predictive sampling. For target column j at sweep t, let $\mathcal{N}_{j,t}$ be its predictor indices. Only rows with $R_{ij} = 1$ train this target's regression; the predictors on those

rows may contain current imputations. A Bayesian ridge model fits a Gaussian linear conditional with a Gaussian shrinkage prior on its coefficients, estimating the noise and prior precisions from the training data. Denote its predictive mean and standard deviation for row i by $\widehat{\mu}_{ij,t}$ and $\widehat{s}_{ij,t}$. The update is

$$\widetilde{Z}_{ij} \leftarrow \begin{cases} Z_{ij}, & R_{ij} = 1, \\ \widehat{\mu}_{ij,t} + \widehat{s}_{ij,t}\xi_{ij,t}, \quad \xi_{ij,t} \sim \mathcal{N}(0,1), & R_{ij} = 0. \end{cases} \tag{23}$$

Here both predictions use the current $\widetilde{Z}_{i,\mathcal{N}_{j,t}}$. A zero predictive standard deviation gives the predictive mean. Equation 23 describes predictive draws; the implementation does not separately sample a regression-coefficient vector at each update. This introduces stochastic conditional variation into missing entries without resampling the observed entries.

How K-MICE chooses predictors. Write K_{imp} for the predictor budget, distinct from the bridge cap K and the cross-fitting count K_{cf}. Full MICE uses all $p-1$ other columns. When the budget covers every other column, K-MICE coincides with full MICE; the effect stage and the HPP campaign use it this way. When $1 \le K_{\mathrm{imp}} < p$, the restricted implementation draws K_{imp} distinct predictor indices for each target update. The weights are computed from the initial mean-filled table, not from the evolving imputed values. With $\varepsilon = 10^{-6}$, define

$$a_{\ell j} = \begin{cases} 0, & \ell = j, \\ \max\{\varepsilon, \min(1, |\operatorname{corr}(\widetilde{Z}^{(0)}_{\cdot\ell}, \widetilde{Z}^{(0)}_{\cdot j})|)\}, & \ell \neq j, \end{cases} \qquad w_{\ell j} = \frac{a_{\ell j}}{\sum_{u \neq j} a_{uj}}. \tag{24}$$

An undefined correlation is assigned the floor ε. At each visit to j, sampling is without replacement and uses probabilities $w_{\cdot j}$; these are sampling weights, not the marginal inclusion probabilities of the resulting subset. Neighbor sets are redrawn across sweeps while their probability reference remains fixed. This is not a deterministic top-K_{imp} rule. If no budget is specified, or it is at least p, all other columns are used. Restricting the predictors changes the fitted conditional model and need not preserve the statistical quality of full MICE.

Mixed-type output and one retained completion. All conditional fits above are numerical Gaussian regressions, including those for integer-coded variables. After the last sweep, and not between individual updates, binary imputations are rounded and clipped to $\{0,1\}$ and an L_j-level categorical imputation to $\{0,\dots,L_j-1\}$. Continuous columns are transformed back to their original units. In the Step 1 wrapper, only originally missing cells receive these completed values; observed cells are protected, with exact preservation checked in standardized space and a 10^{-8} round-trip tolerance on the original scale. Non-finite output is not saved. This projection enforces legal codes; it does not turn a Gaussian regression into a logistic or multinomial conditional, and nominal-category coding can affect the imputation.

Although MICE conventionally supplies multiple completed data sets, this pipeline keeps one stochastic completion for discovery. There is no pooling across independent completions and no Rubin variance correction. Correctly specified missingness and conditional-model assumptions are still needed to recover a full-data law; posterior sampling alone provides neither this recovery nor valid causal identification. The missing-data modes of Section 3.5 still apply.

Memory reductions in our implementation. The standard restricted backend stores the $p \times p$ absolute-correlation matrix. Our `streaming_replay` option instead retains a centered, immutable copy of $\widetilde{Z}^{(0)}$ and its column norms. It reconstructs one probability column of equation 24 when a target needs it and uses the inherited random-number call location and weighted sampling rule. This removes the retained $O(p^2)$ correlation matrix, but does not remove the input table or the fitted conditional models. In particular, retaining all Bayesian ridge models can still require $O(T_{\mathrm{imp}} p K_{\mathrm{imp}}^2)$ covariance storage.

For a one-shot `fit_transform` call, the additional `fit_transform_only` policy discards each fitted regression after its current prediction and write-back. Target and neighbor indices are retained as the update trace. With both options enabled and $K_{\mathrm{imp}} < p$, the implementation's retained data, trace and fitting workspace have the bound

$$O\big(np + T_{\mathrm{imp}} p K_{\mathrm{imp}} + nK_{\mathrm{imp}} + K_{\mathrm{imp}}^2\big)\,. \tag{25}$$

This is a storage claim for the imputer, not for the subsequent discovery algorithm. Reconstructing all probability columns every sweep still entails $\Theta(T_{\mathrm{imp}} np^2)$ correlation work. The streaming op-

Algorithm 2 One-shot discovery completion by MICE / K-MICE

Require: $D \in (\mathbb{R} \cup \{\mathrm{NA}\})^{n\times p}$, $R \in \{0,1\}^{n\times p}$, types; T_{imp}, $K_{\mathrm{imp}} \in \mathbb{N}_{>0} \cup \{\perp\}$, seed ζ; streaming/release flags $b_s, b_r \in \{0,1\}$.

Ensure: One completion $\widetilde{D}$ and update trace $\mathcal{T}$; $b_r = 1$ excludes later transforms.

$O_j \leftarrow \{i : R_{ij} = 1\}$, $U_j \leftarrow \{i : R_{ij} = 0\}$; require $|O_j| > 0$ for all j.
$\mathcal{S} \leftarrow \mathrm{FitScale}(D, R)$; $Z \leftarrow \mathcal{S}(D)$; $\mathcal{T} \leftarrow ()$.
$\widetilde{Z}_{O_j,j} \leftarrow Z_{O_j,j}$; $\widetilde{Z}_{U_j,j} \leftarrow |O_j|^{-1} \sum_{i\in O_j} Z_{ij}$ for all j.
$\widetilde{Z}^{(0)} \leftarrow \widetilde{Z}$ (fixed copy); $\sigma \leftarrow \mathrm{argsort}_j\, |U_j|$; $\mathrm{rng} \leftarrow \mathrm{RNG}(\zeta)$.
$r \leftarrow \mathbf{1}\{K_{\mathrm{imp}} \in \{1, \ldots, p-1\}\}$; $T_* \leftarrow T_{\mathrm{imp}} \mathbf{1}\{p > 1\}$.
If $r = 1$ and $b_s = 0$, cache $W = (w_{\ell j})$ from equation 24.
for $t = 1, \ldots, T_*$ **do**
  for $j \in \sigma$ **do**
    If $r = 1$, read $w_{\cdot j}$ from W, or reconstruct it from $\widetilde{Z}^{(0)}$ if $b_s = 1$.
    $\mathcal{N}_{j,t} \leftarrow \begin{cases} \mathrm{WOR}_{K_{\mathrm{imp}}}(w_{\cdot j}; \mathrm{rng}), & r = 1, \\ \{1, \ldots, p\} \setminus \{j\}, & r = 0. \end{cases}$
    $\mathcal{B}_{j,t} \leftarrow \mathrm{BayesianRidge.fit}(\widetilde{Z}_{O_j,\mathcal{N}_{j,t}}, Z_{O_j,j})$.
    if $|U_j| > 0$ **then**
      $(\widehat{\mu}_{ij,t}, \widehat{s}_{ij,t})_{i\in U_j} \leftarrow \mathrm{Predict}(\mathcal{B}_{j,t}, \widetilde{Z}_{U_j,\mathcal{N}_{j,t}})$.
      $\widetilde{Z}_{ij} \leftarrow \begin{cases} \widehat{\mu}_{ij,t} + \widehat{s}_{ij,t}\xi_{ij,t}, & \xi_{ij,t} \sim \mathcal{N}(0,1), \quad \widehat{s}_{ij,t} > 0, \\ \widehat{\mu}_{ij,t}, & \widehat{s}_{ij,t} = 0, \end{cases} \quad i \in U_j.$
    end if
    $\mathcal{B}_{j,t} \leftarrow \perp$ **if** $b_r = 1$.
    $\mathcal{T} \leftarrow \mathcal{T} \,\|\, [(j, \mathcal{N}_{j,t}, \mathcal{B}_{j,t})]$.
  end for
end for
$\widetilde{Z}_{U_j,j} \leftarrow \Pi_j(\widetilde{Z}_{U_j,j})$ for all j; $\widetilde{D} \leftarrow \mathcal{S}^{-1}(\widetilde{Z})$.
Validate row order, $\widetilde{D} \in \mathbb{R}^{n\times p}$ and $\max_{R_{ij}=1} |\widetilde{D}_{ij} - D_{ij}| \le 10^{-8}$.
return $(\widetilde{D}, \mathcal{T}, \mathrm{configuration})$.

$\mathcal{S}$: observed-entry continuous scaling; Π_j: identity for continuous variables, round-and-clip for discrete variables. WOR: weighted sampling without replacement; $\|$: trace append. All draws use rng; $\perp$ denotes an unset budget or a released model.

tion trades recomputation for storage; no linear-time or wall-clock speedup claim follows. Cache release explicitly disallows later transformation of new rows, so it cannot be used for the training-fit/validation-transform scheme of Appendix C, which requires retaining the fitted models. The adapter checks compatibility with scikit-learn 1.4.2 and the private methods it overrides; compatibility with another version is not assumed.

Why use chained conditionals rather than a latent reconstruction model? The choice is motivated by control of the completion: each conditional exposes its predictor set, its observed training rows and its predictive draw, and a neighbor budget gives direct control of regression size. It also avoids selecting and training a separate encoder–decoder architecture. These are reasons for choosing this implementation, not evidence that MICE dominates other imputers. VAE completion, as described in Li et al. (2025), is an alternative. VAEs can sample both latent states and observation distributions, and the VAE backend retained in our code does so; therefore the change is not justified by claiming that VAEs are inherently deterministic or must suppress variance. Nor do the present experiments isolate an effect-accuracy gain caused by replacing VAE completion with MICE. Gaussian conditional misspecification, categorical projection and nonignorable missingness remain possible sources of bias.

Defaults, implementation checks and experimental scope. The Step 1 backend defaults to Bayesian ridge, posterior sampling, ten sweeps, seed zero and no predictor restriction. Both the streaming backend and model-cache release are opt-in: the defaults remain `dense_sklearn` and

`retain_transform_state`. A K-MICE run must therefore report its predictor budget and enabled options rather than infer them from the backend's availability. Development replay checks at $p \in \{100, 500, 2000\}$, $n = 600$, $K_{\mathrm{imp}} = 16$ and $T_{\mathrm{imp}} = 1$ compared streaming and dense implementations, and separate ten-sweep checks assessed cache release. The tested outputs, neighbor schedules and final random states agreed exactly. This is evidence on the tested configurations, not a proof of bitwise equivalence for every input, floating-point backend or library version. These checks are implementation validation, not the FGES discovery benchmark or an imputation-quality comparison. K-MICE completes the inputs of the discovery benchmark of Table 1, with the budget reported in Appendix G.2; the effect-only benchmark and the main end-to-end campaigns use complete data, and the missing-data check of Appendix I.5 uses it for discovery.

G.2 Configuration used in the discovery benchmark

Section 3.2 states the discovery stage in four steps. Here we give the settings, diagnostics and scored outputs used for Table 1.

Input and the information boundary. The stage receives a table over an exposure, an outcome, a registered list of candidate mediators and the remaining measured variables, each with a declared type of continuous, binary or categorical. Missing entries are completed by K-MICE (Algorithm 2), which returns every column declared binary or categorical on its observed levels, so a discrete variable is still discrete for the score of equation 27. The completion uses a predictor budget $K_{\mathrm{imp}} = 64$ and otherwise the defaults of Appendix G.1, namely Bayesian ridge with posterior sampling, $T_{\mathrm{imp}} = 10$ sweeps, seed zero, the dense backend and retained transform state; it runs once per replicate and its output is shared by the four methods. The screen is called with its ground-truth arguments left empty, which are the true adjacency, the true path edges and the true path list, so no function of the true graph can reach it. The role information the stage does consume is the index of the exposure, the index of the outcome and the list of candidate mediator indices. Regenerating the screen for a cell recomputes its edge count and compares it against the published count for that cell, and a mismatch is fatal rather than a warning.

The candidate screen. The screen implements four sources whose union forms the candidate set, and two of them, the correlation and role sources, are active in this configuration. The correlation source computes the absolute Pearson correlation between every pair of columns, sets values that are not finite to zero, clears the diagonal, and retains for each variable the 32 strongest partners. The mutual-information source is switched off and contributes an empty matrix. The role source admits every ordered pair drawn from the four registered role pair types, which are exposure to candidate mediator, candidate mediator to outcome, exposure to outcome, and candidate mediator to candidate mediator. The exploratory source is given a fraction of zero and an edge budget of zero, so it contributes nothing. The union is taken, self-loops are cleared, and equation 26 symmetrises the result into the searchable domain. Selection within a variable is by descending score and the implementation audits the boundary rather than asserting that it is unambiguous. It counts exact ties at the retained rank, counts scores within a tolerance of that rank, records the smallest positive gap at the boundary and the variable it occurred on, and counts the cases in which the retained order was not determined by the scores alone.

Writing $\mathcal{E}_{\mathrm{scr}}$ for the retained ordered pairs, the undirected search domain is

$$\mathcal{E}_{\mathrm{allow}} = \big\{\{i, j\} : i < j,\ (i, j) \in \mathcal{E}_{\mathrm{scr}} \ \vee\ (j, i) \in \mathcal{E}_{\mathrm{scr}}\big\}. \tag{26}$$

Membership permits an edge without forcing its presence or direction.

The restricted search. The searchable domain is written into an undirected graph over the same variables and handed to a greedy equivalence search through its bound-graph setting, which restricts every addition and removal to that domain. The objective is a Bayesian information criterion of degenerate-Gaussian form. Its structure prior is zero and its singularity tolerance is zero, and the penalty discount together with the maximum degree are the two settings that vary between runs. Faithfulness is not assumed and the first step is not symmetrised. The role tier is supplied as forbidden edges, from the outcome to the exposure, from the outcome to each candidate mediator, and from each candidate mediator to the exposure. Each arm runs in its own process with an explicit heap, so an exhausted heap or a failure is recorded against that arm alone and is separated in the record from an arm that returned.

Let $\mathcal{K}_{\mathrm{role}}$ encode the forbidden edges of the registered tiers $X \prec m \prec Y$ for $m \in \mathcal{M}$, and let λ collect the maximum degree and penalty discount fixed for each run. The search output is

$$\widehat{H} = \mathrm{FGES}\big(\widetilde{D};\ \mathcal{E}_{\mathrm{allow}},\ \mathcal{K}_{\mathrm{role}},\ \lambda\big). \tag{27}$$

The reference FGES implementation is unchanged, and $\widehat{H}$ is not a certified global optimum.

Reading the returned graph. The search returns an endpoint matrix. A pair carrying a tail at one end and an arrow at the other is read as a directed edge, and a pair that is present in any other configuration is read as an adjacency whose direction was not committed. A return that carries no uncommitted adjacency at all is mapped onto the completed partially directed graph of its equivalence class before scoring and the substitution is recorded, so a search that happened to return a single fully oriented member is not credited with directions its class cannot identify. An unresolved pair is therefore emitted rather than guessed, which is the convention the orientation columns of Appendix F are defined against.

Role-edge stages. No orientation repair or role-edge recovery acts on the graph that Table 1 scores or on the end-to-end graph family. The earlier discovery lane that supplies the auxiliary weights of Appendix G.3 applies its own orientation and role-edge repairs before its post-repair weights are read, and neither weight version changes a graph.

Which output the table scores. The accuracy columns of Table 1 are scored on the graph the search returns, read under the endpoint convention above, before any graph family is built. Pairs searched is the size of the searchable domain as a share of all pairs, and wall-time and memory measurements come from execution records, with their measurement conventions specified in Appendix F.

The bound on the searchable domain and what it does not cover. The correlation source contributes at most the per-variable budget times the number of variables. The role source contributes at most the number of ordered pairs its four types admit, which for a single exposure and a single outcome is the number of candidate mediators squared, plus that number, plus one. The two remaining sources contribute nothing here. The searchable domain is therefore linear in the number of variables up to a term in the size of the registered mediator list, against a pair space that is quadratic. Two things that bound does not cover are worth stating. The correlation backend these runs use forms the dense correlation matrix over all variables, so the screen that produces a linear domain is itself quadratic in memory, and a streaming backend that avoids this exists but was not used. And the bound constrains how many relations are examined, not how long the search takes or how much memory it holds, both of which Section 4.1 measures directly.

What the stage hands on. The stage writes one partially directed graph over every variable, the set of committed directed edges, the set of uncommitted adjacencies, the role subgraph, the size of the searchable domain with its share of the pair space, the terminal state of the arm, its wall time and its memory reading. In the end-to-end configuration this object becomes the graph family of Section 3.3, built as described in Appendix G.3.

G.3 Settings used end to end

Knowledge and search domain. The end-to-end campaigns give the search a complete public order of all variables in five tiers, public baseline covariates, the exposure, the registered mediators, the public outcomes and post-target background, with no order inside a tier. Every column must belong to one tier, the tiers are supplied to the searcher as native background knowledge, and the implementation reads the tier lists and the full forbidden-edge relation back before the search runs. The screened domain of equation 26 is enlarged by every pair from a public baseline covariate to the exposure, a registered mediator or the outcome, and none of these pairs is required. Possible instruments among the baseline covariates are treated like any other baseline covariate. The search uses the degenerate-Gaussian score with penalty discount 4 and maximum degree 32, with the remaining settings of Appendix G.2.

The graph family. Each of the $S{=}30$ initial graphs is a consistent extension of the search output. A draw keeps the input skeleton, the input directed edges and exactly the input unshielded colliders,

satisfies every forbidden edge, and is built by removing at each step a sink chosen uniformly from the sorted set of admissible sinks (Wienöbst et al., 2021). Draws are taken with replacement and a repeated graph stays a separate member, so the law is not uniform over distinct graphs, not a posterior and not a bootstrap. Each distinct initial graph is then refined. For every registered mediator and outcome in turn, the refinement considers replacing its parent set by one that removes at most two parents and adds at most two, drawn from the screened search domain less the forbidden edges, and that create no cycle. It keeps the change with the largest gain in the local degenerate-Gaussian score when that gain exceeds 10^{-8}, respects the degree cap of 32, and repeats for at most five passes. The refinement is deterministic, reads no truth and no effect data, and keeps the thirty draw multiplicities. A refined graph is in general not an extension of the search output. After refinement every one of the thirty graphs is checked for a cycle and for a directed path from a later public tier to an earlier one. If any graph fails, no set is issued for the replicate and all thirty draws stay in its record. The directed edges the search returns pass the same check before any draw is taken, and a failure there stops the replicate. All forty replicates of the $p{=}500$ end-to-end campaign passed both checks.

The auxiliary branch. For each replicate the auxiliary branch computes two versions of a candidate-edge weight matrix from the standardised discovery data, using the correlation and role screen of Appendix G.2 before the baseline pairs are added. The pre-repair weight of a candidate edge into a variable is the absolute coefficient of an ordinary least-squares regression of that variable on all of its screened candidates, divided by the sum of these absolute coefficients. The bridge rebuilds these weights independently and stops if they disagree with the branch's own. The operating threshold comes from one copy of the data in which every column is permuted independently, the screen is rebuilt and the weights are refitted. With the larger of its two directional weights as the score of a pair, the threshold is the larger of the 0.999 quantile of the non-zero permuted scores and a tail guard, the smallest observed score retained by a Benjamini–Hochberg step at level 0.05 on empirical permutation p-values, each the share of pooled non-zero permuted scores at or above the observed score, with one added to numerator and denominator. The guard is descriptive and controls neither the false discovery rate nor the family-wise error; in every replicate of the $p{=}500$ end-to-end campaign the quantile was the larger term. The post-repair version applies three steps to a copy of the pre-repair weights. A non-Gaussian orientation step visits each pair of continuous variables screened in both directions whose larger weight clears the threshold; when an additive-noise comparison of the two directions is confident and at least one of the two variables has a non-Gaussian residual, it places the larger weight on the preferred direction and zeroes the other. A direction is scored by the distance correlation between its putative cause and the residual of its effect, regressed on that cause and on the five screened candidate parents of the pair with the largest weights; the comparison counts as confident when the larger of the two scores is at least 1.5 times the smaller and the smaller is below 0.15, and a residual counts as non-Gaussian when a D'Agostino–Pearson test on that variable, regressed on the same parents, returns $p < 10^{-3}$. A role step, for each pair among the exposure, the mediators and the outcome with a single role-permitted direction that the screen admits, places the larger of the two weights on that direction and zeroes the other, leaving pairs of mediators unchanged. A final step adds exposure, mediator and outcome edges chosen by role-specific regressions, with a Bonferroni-corrected t test at level 0.05 within each edge type (0.01 for mediator-to-mediator edges), a coefficient floor of 0.02 and at most two parents per mediator; each added edge takes the largest post-repair weight and its reverse is zeroed. The branch stops outright, if a repair alters the pre-repair weights, a permuted copy yields no non-zero score, the guard retains no pair or the threshold is not finite and positive. Paths that clear the stability threshold are ranked on the pre-repair weights, the fallback ratios of Appendix B are formed from the post-repair weights and the threshold, and neither version adds, removes or orients an edge of any graph.

Path selection. A registered path $X \to M \to Y$ is eligible when both of its edges occur together in at least six of the thirty graphs, which is $\widehat{\mathrm{PSS}} \geq 0.2$. Eligible paths are ranked by the geometric mean of their two pre-repair auxiliary edge weights, then by the smaller of the two, then by registry order, and the first is selected. When no registered path is eligible the bridge issues no set and records the reason.

H SUPPLEMENTARY PANELS FOR EFFECT ESTIMATION

These panels accompany Table 2. The first separates the cost of a flexible commitment from its benefit, and the second compares a linear and a flexible outcome estimator across the nonlinear share on one family. Every quantity uses the protocol of Appendix F.

H.1 COMMITMENT COST AGAINST MISSPECIFICATION EXPOSURE

Figure 5 shows every method of Table 2 except the two mediation-specific estimators, which are compared below, in two panels: the cost it pays where a linear form is already correct, and its exposure on the worst nonlinear family. Both are excess $100L_{\mathrm{rm}}$ over the oracle-form reference, formed within replicate and within family; intervals are clustered paired bootstrap 95% intervals over base replicates with the worst family re-selected inside every resample. Linear-DML is a partially linear estimator: the outcome and the terms x, m and xm are residualised on C_π by cross-fitted networks, and the effects are read from a least-squares fit on those residuals.

The right panel carries the result. The three estimators that are linear in the mediator, Parametric g, PL-DML and Linear-DML, reach between 32 and 65 points on their worst nonlinear family, against at most 8.8 for any method in the figure that is flexible in the mediator and under one point for CHARM under both exposure profiles. The left panel gives the cost of that protection. CHARM pays 0.87 points on the linear family under continuous exposure and 0.18 under binary exposure, where the plain `MLP` baseline pays 9.27 and 4.46. Pooled over both exposure profiles and paired within replicate, the plain baseline exceeds CHARM by $+6.34$ points of $100L_{\mathrm{rm}}$ on the linear family (clustered bootstrap 95% interval $[+4.19, +8.52]$, positive in $10/10$ replicates) and by $+4.66$ on the nonlinear families ($[+3.10, +6.44]$, $10/10$). The two share a base learner but differ in their training settings as well as in the additive form, so the gap is not attributable to the additive form alone. A flexible commitment is cheap here, but not automatically so.

Against the two flexible comparators the paired differences are small and do not separate from zero. Formed within replicate as comparator minus CHARM on the equal-family mean, in points of $100L_{\mathrm{rm}}$ and with the same clustered bootstrap, CV-selected exceeds CHARM by $+1.47$ $[-0.82, +3.72]$ under continuous and $+0.21$ $[-0.11, +0.55]$ under binary exposure, being higher in $7/10$ and $6/10$ replicates, and Spline exceeds it by $+1.81$ $[-0.98, +4.62]$ and $+0.27$ $[-0.09, +0.63]$, higher in $7/10$ replicates under each. On the worst family both comparators exceed CHARM by $+2.27$ under continuous and $+0.13$ under binary exposure, again with intervals that include zero. Family by family, CV-selected has the lower point estimate on the linear family under both profiles and on two nonlinear families under binary exposure; the largest such gap is 0.85 points on the linear family under continuous exposure, where it selects ridge, with a 95% t interval of $[-1.74, +0.03]$.

The two mediation-specific estimators trail CHARM clearly. Paired within replicate, mediation DML exceeds it by $+7.56$ points of $100L_{\mathrm{rm}}$ $[+3.52, +12.16]$ on the equal-family mean under binary exposure, higher in $9/10$ replicates, and by $+9.98$ $[+4.82, +16.54]$ on the worst family; GPS weighting exceeds it by $+19.72$ $[+13.10, +27.42]$ under continuous exposure, higher in $10/10$, and by $+38.22$ $[+27.92, +48.28]$ on the worst family. The DML gap is carried by the indirect effect and is variance rather than bias: its sign-aligned mean NIE error lies between -0.006 and $+0.021$ in every family, and its direct-effect error matches that of CHARM. Trimming rows with propensities outside $[0.05, 0.95]$ instead of clipping removes a median of 23.5% of rows and raises its mean L_{rm} to 26.02%.

H.2 LINEAR VERSUS FLEXIBLE OUTCOME ESTIMATOR ACROSS THE NONLINEAR SHARE

Design. Table 2 fixes the nonlinear share at η=0 and η=0.95. This comparison fills in the interval on the sine family, holding the shape fixed and sweeping $\eta \in \{0, 0.25, 0.5, 0.75, 0.95\}$, a cut orthogonal to the panel, which sweeps shape at fixed strength. It pairs CHARM with Linear-DML, the partially linear estimator of Appendix H.1, on the same data. The two differ in the outcome form, Linear-DML being linear in x and m with an xm term, and in how the adjustment channel is fitted, so D compares the two estimators as wholes rather than isolating the mediator channel. We plot $D(\eta) = 100\{L_{\mathrm{rm}}^{\text{Linear-DML}} - L_{\mathrm{rm}}^{\text{CHARM}}\}$, so $D > 0$ means the flexible estimator is the safer one. The η=0.95 rung is the Sin cell of Table 2, and the other rungs use independent seeds.

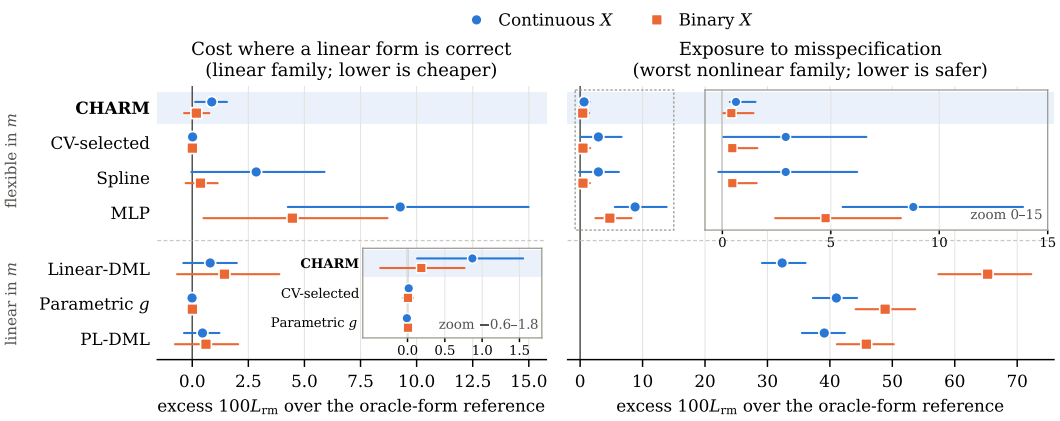


Figure 5: Commitment cost on the linear family (left) and misspecification exposure on the worst nonlinear family (right), both as excess $100L_{\mathrm{rm}}$ over the oracle-form reference, for each method and exposure profile; lower is better in both panels. Bars are clustered paired bootstrap 95% intervals, with the worst family re-selected inside every resample. Linear-DML is the partially linear estimator described in the text, and the insets zoom CHARM, CV-selected and Parametric g near zero on the left and the four methods flexible in the mediator to 0–15 on the right. On the linear family the oracle-form reference is the correctly specified linear model, so Parametric g, which fits that model with cross-fitting, and CV-selected, which selects a ridge model there in all twenty cells, sit close to zero in this experiment; the left panel is read together with the right.

Result. At η=0 the two estimators are indistinguishable ($D = 0.13$, 95% t interval $[-0.95, 1.21]$ continuous; 0.17, $[-2.20, 2.54]$ binary), so the flexible estimator carries no detectable cost where the linear form is correct. The difference is small at η=0.25 and grows monotonically from there: D reaches 11.01 $[5.71, 16.30]$ at η=0.50, 22.18 $[17.32, 27.03]$ at η=0.75 and 31.73 $[27.00, 36.46]$ at η=0.95 under continuous exposure, and 17.41 $[13.11, 21.71]$ at η=0.95 under binary exposure. No rung has D significantly below zero, so the sweep locates where the linear estimator becomes unsafe but does not establish a regime in which it is preferable.

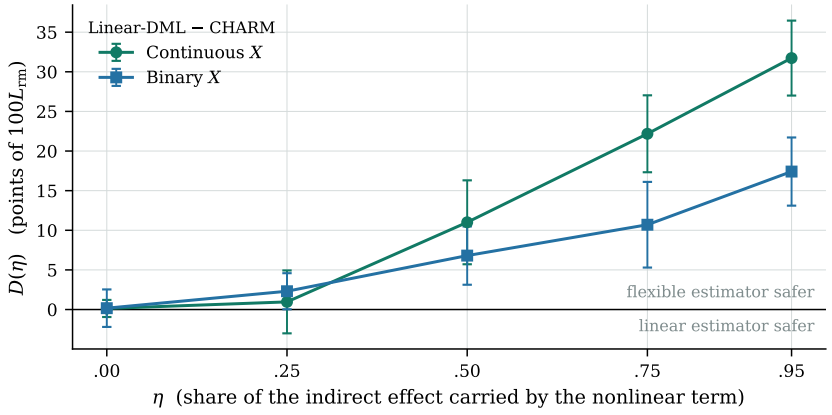


Figure 6: Linear versus flexible outcome estimator as a function of the nonlinear share η on the sine family: $D(\eta) = 100\{L_{\mathrm{rm}}^{\text{Linear-DML}} - L_{\mathrm{rm}}^{\text{CHARM}}\}$ with 95% paired t intervals over R=10 replicates. At η=0 the two estimators are indistinguishable, and the cost of the linear estimator grows monotonically to 31.7 points of $100\,L_{\mathrm{rm}}$ under continuous exposure. **No rung is significantly negative**, so the sweep does not establish a regime in which the linear estimator is preferable.

Figure 6 plots only the paired difference D, and Table 6 gives the two arms it is built from. D is formed within replicate before being averaged, so no comparison crosses datasets, and its interval is the 95% t interval of that paired difference over the ten replicates of a rung. Across the sweep

the CHARM column stays between 5.66 and 6.92 under continuous exposure and between 8.40 and 9.83 under binary exposure, while the Linear-DML column rises from 7.05 to 37.39 and from 10.00 to 25.81.

Table 6: Per-rung values for the η sweep of Figure 6, in points of $100\,L_{\mathrm{rm}}$ (sine family, p=100, n=10^4, complete data). Every rung is complete at R=10; no replicate is dropped and no rung is short. D is the within-replicate paired difference with its 95% t interval, and Linear-DML is the partially linear estimator of Appendix H.1.

Exposure	η	R	Linear-DML	CHARM	D [95% t]
Continuous	0.00	10	7.05	6.92	0.13 [−0.95, 1.21]
	0.25	10	7.09	6.12	0.96 [−3.01, 4.93]
	0.50	10	16.89	5.89	11.01 [5.71, 16.30]
	0.75	10	28.19	6.01	22.18 [17.32, 27.03]
	0.95	10	37.39	5.66	31.73 [27.00, 36.46]
Binary	0.00	10	10.00	9.83	0.17 [−2.20, 2.54]
	0.25	10	11.68	9.37	2.31 [0.03, 4.58]
	0.50	10	15.81	9.00	6.81 [3.13, 10.48]
	0.75	10	20.50	9.80	10.70 [5.30, 16.10]
	0.95	10	25.81	8.40	17.41 [13.11, 21.71]

I THE END-TO-END BRIDGE IN FULL

Table 3 compresses the bridge experiment to the quantities the main claim needs. This appendix gives the full compiled-set error panel in Table 7 and the matched comparison arms of Appendix B.6 in Table 8, at all three scales. Within a replicate the three arms share the complete data, the discovery output, the learned path, the split, the effect-stage settings and the random-number streams, and differ in the source of the adjustment set; the compiled and supplied arms also fitted the same outcome and mediator classes in every replicate, and the unadjusted arm differs in its outcome model in one replicate, as recorded below. The protocol behind the campaign is in Appendix F. Appendices I.5 and I.6 report the missing-data check and the stress tests summarised in Table 3.

I.1 FULL COMPILED-SET EFFECT-ERROR PANEL

Table 7: Complete effect-error metrics for the compiled-set arm at each scale. Bias is the mean signed error $\widehat{\theta} - \theta^{\star}$, while MAE, RMSE and MaxAE are computed after taking errors within replicate, so opposite signs cannot cancel in the latter three columns. Sign is the fraction of replicates whose estimated effect has the correct sign, and the value after $\pm$ is the standard error of the replicate mean. Every planned replicate enters the table.

p	Effect	Bias	MAE	RMSE	MaxAE	Sign
500	NIE	+0.0003	0.0080 ± 0.0009	0.0096	0.0222	1.00
	NDE	−0.0014	0.0102 ± 0.0012	0.0125	0.0306	1.00
	TE	−0.0011	0.0104 ± 0.0014	0.0135	0.0432	1.00
5000	NIE	+0.0043	0.0043 ± 0.0009	0.0046	0.0060	1.00
	NDE	+0.0051	0.0051 ± 0.0020	0.0065	0.0116	1.00
	TE	+0.0094	0.0094 ± 0.0019	0.0101	0.0164	1.00
10000	NIE	+0.0026	0.0029 ± 0.0012	0.0037	0.0066	1.00
	NDE	+0.0081	0.0081 ± 0.0026	0.0096	0.0141	1.00
	TE	+0.0107	0.0107 ± 0.0032	0.0125	0.0196	1.00

I.2 ADJUSTMENT-SET COMPARISON ARMS

The compiled- and supplied-set arms fitted the base class in every replicate, with a linear conditional mean for the mediator throughout, and the pre-registered interaction diagnostic of Appendix C fired in neither arm. In the unadjusted arm it fired in replicate 35, which therefore carries the declared varying-coefficient extension and an estimate labelled supported with a warning. That diagnostic reads the effect-stage data before outer-fold fitting, so it is the one step that can add the declared extension to the base class.

Table 8: Matched three-arm comparison for the end-to-end bridge at each scale. Compiled selects one learned set with no user-specified C_π, Supplied uses the deterministic oracle-valid set on the same learned path and is the paired accuracy anchor, and No-C fixes $C_\pi = \varnothing$ as a negative control. Entries are NIE/NDE/TE MAE and the combined L_{rm} in percentage points, with the standard error of the replicate mean after $\pm$. The compiled and supplied arms use the same adjustment set in every replicate and return identical estimates. All three arms have the correct sign on every contrast in every replicate, which is why sign alone is not an adequate accuracy metric.

p	Adjustment arm	NIE MAE	NDE MAE	TE MAE	L_{rm}
500	Compiled	0.0080	0.0102	0.0104	4.54 ± 0.40
	Supplied	0.0080	0.0102	0.0104	4.54 ± 0.40
	No-C	0.0520	0.1320	0.1045	46.01 ± 1.15
5000	Compiled	0.0043	0.0051	0.0094	2.36 ± 0.47
	Supplied	0.0043	0.0051	0.0094	2.36 ± 0.47
	No-C	0.0598	0.1410	0.0921	50.20 ± 3.22
10000	Compiled	0.0029	0.0081	0.0107	2.76 ± 0.82
	Supplied	0.0029	0.0081	0.0107	2.76 ± 0.82
	No-C	0.0462	0.1405	0.1192	46.68 ± 3.27

I.3 POST-REVEAL ADJUSTMENT-SET DIAGNOSTICS

Table 9: Adjustment-set diagnostics read after truth reveal, with the selector branch that produced each set and the wall time. Graph-admissible is the joint gate of Table 3. Oracle-exact counts the replicates whose compiled set equals the registered two-variable reference set, which is not B_{GT}, and at $p{=}500$ it coincides with the admissibility gate in every replicate. Descendant-free counts sets containing no true descendant of the exposure. Refused counts paths that received no set. One-sided panel and Fallback count the replicates in which the one-sided-panel condition fired and in which the fallback supplied the set. Wall is the median time of one replicate from start to exit, in hours.

		Adjustment set				Selector branch			Resources
p	R	Graph-adm.	Oracle-exact	Desc.-free	$\lvert\widehat{C}_\pi\rvert$	One-sided panel	Fallback	Refused	Wall (h)
500	40	$40/40$	$40/40$	$40/40$	2	$0/40$	$0/40$	0	0.27
5000	5	$5/5$	$5/5$	$5/5$	2	$0/5$	$0/5$	0	9.25
10000	5	$5/5$	$5/5$	$5/5$	2	$0/5$	$0/5$	0	63.09

Table 10: The $p{=}500$ assessment by registered stratum, whose structures Figure 4 shows. Each stratum was designed before the campaign and contributes ten replicates to the fixed denominator of forty. Three of the four place a structure adversarial to the compiler in the generator, and every stratum passes $10/10$. The total minimum, $37/40$, is the smallest count whose two-sided 95% Wilson lower bound clears the registered 0.80.

Registered stratum	Graph-admissible	Registered minimum
Ordinary	$10/10$	$9/10$
Near-instrument	$10/10$	$9/10$
Collider descendant	$10/10$	$9/10$
Strong exposure descendant	$10/10$	$9/10$
Total	$40/40$	$37/40$

I.4 EVERY REPLICATE OF THE END-TO-END CAMPAIGNS

Table 11 lists every replicate behind the rows of Table 3. It is included because statements in the main text concern individual replicates rather than an average, and the admissibility gate and the agreement between the two arms can be checked one row at a time.

Table 11: Every replicate of the three end-to-end campaigns. *Str.* is the registered stratum, ordinary, near-instrument, collider-descendant or strong-exposure-descendant. *Adm.* is the joint admissibility gate of Table 3. The three effect columns are the compiled arm. The last column is $=$ where the supplied arm returned the same three estimates to floating-point equality. Every replicate selects the exact target path, so that column is omitted. The two larger campaigns were not stratified, so every replicate there is ordinary.

p	Rep	Str.	Adm.	NIE	NDE	TE	Sup.
500	1	ord	✓	+0.2067	+0.2009	+0.4076	=
	2	n-i	✓	+0.1992	+0.1857	+0.3849	=
	3	c-d	✓	+0.2131	+0.1694	+0.3826	=
	4	s-d	✓	−0.2004	−0.1978	−0.3982	=
	5	ord	✓	+0.2075	+0.1939	+0.4014	=
	6	n-i	✓	+0.1936	+0.1835	+0.3771	=
	7	c-d	✓	+0.1957	+0.2101	+0.4058	=
	8	s-d	✓	+0.1880	+0.2182	+0.4062	=
	9	ord	✓	−0.2038	−0.2188	−0.4226	=
	10	n-i	✓	−0.2169	−0.2002	−0.4172	=
	11	c-d	✓	−0.1913	−0.2105	−0.4019	=
	12	s-d	✓	−0.1934	−0.2075	−0.4010	=
	13	ord	✓	−0.2125	−0.2093	−0.4218	=
	14	n-i	✓	+0.2030	+0.2105	+0.4135	=
	15	c-d	✓	−0.1976	−0.2068	−0.4044	=
	16	s-d	✓	−0.2002	−0.2056	−0.4058	=
	17	ord	✓	+0.1879	+0.1888	+0.3767	=
	18	n-i	✓	+0.1867	+0.2236	+0.4104	=
	19	c-d	✓	+0.1919	+0.1975	+0.3894	=
	20	s-d	✓	−0.1778	−0.1790	−0.3568	=
	21	ord	✓	−0.1970	−0.2078	−0.4048	=
	22	n-i	✓	−0.2098	−0.1767	−0.3865	=
	23	c-d	✓	+0.2128	+0.1914	+0.4043	=
	24	s-d	✓	−0.2057	−0.1864	−0.3922	=
	25	ord	✓	−0.1795	−0.2112	−0.3907	=
500	26	n-i	✓	+0.1917	+0.2076	+0.3994	=
	27	c-d	✓	+0.1889	+0.2039	+0.3928	=
	28	s-d	✓	−0.1897	−0.1989	−0.3886	=
	29	ord	✓	−0.1901	−0.2014	−0.3915	=
	30	n-i	✓	+0.2137	+0.1832	+0.3969	=
	31	c-d	✓	+0.1925	+0.1927	+0.3852	=
	32	s-d	✓	−0.1878	−0.2094	−0.3972	=
	33	ord	✓	+0.1955	+0.2182	+0.4137	=
	34	n-i	✓	+0.2048	+0.1996	+0.4044	=
	35	c-d	✓	−0.2059	−0.1990	−0.4048	=
	36	s-d	✓	−0.1991	−0.1984	−0.3975	=
	37	ord	✓	+0.2001	+0.1915	+0.3917	=
	38	n-i	✓	+0.2074	+0.1903	+0.3977	=
	39	c-d	✓	+0.1901	+0.1893	+0.3794	=
	40	s-d	✓	−0.1999	−0.1815	−0.3814	=
5000	1	ord	✓	+0.2058	+0.2022	+0.4081	=
	2	ord	✓	+0.2014	+0.2050	+0.4064	=
	3	ord	✓	+0.2034	+0.2066	+0.4100	=
	4	ord	✓	−0.1940	−0.1999	−0.3938	=
	5	ord	✓	+0.2048	+0.2116	+0.4164	=
10000	1	ord	✓	−0.2009	−0.1859	−0.3868	=
	2	ord	✓	−0.1998	−0.1997	−0.3995	=
	3	ord	✓	+0.2025	+0.2048	+0.4073	=
	4	ord	✓	−0.1957	−0.1916	−0.3872	=
	5	ord	✓	+0.2066	+0.2129	+0.4196	=

What the appendix establishes.

1. **The paired supplied-set comparison supports preservation, not superiority.** The compiled and supplied arms use the same adjustment set in every replicate and return identical estimates, so at every scale the paired L_{rm} difference and every paired MAE difference are zero. These are two readings of one paired comparison and not independent evidence, and they give no evidence of an advantage over a supplied valid set.
2. **Using an adjustment set materially changes the estimates.** Removing C raises L_{rm} from 4.54 for the compiled arm to 46.01 at $p{=}500$, a 10.1-fold increase, and from 2.36 to 50.20 and from 2.76 to 46.68 at the two larger scales, while the No-C arm keeps the correct sign on every contrast, so direction alone can conceal large magnitude errors.
3. **The learned policy is exact on this benchmark.** The compiled set equals the registered two-variable reference set in all forty replicates, and in all five at each larger scale, which is what makes the paired arms identical. This is a property of a benchmark whose registered reference sets all have two members, matching the registered set size, and no replicate of these campaigns required the bridge to issue no set, a case the stress tests of Appendix I.6 cover.
4. **True-graph admissibility holds in every replicate.** After truth reveal all forty sets pass the joint admissibility check and all forty are free of true exposure descendants. The lower end of the two-sided 95% Wilson interval is 0.912, and Table 10 shows every stratum at $10/10$. The two larger campaigns are $5/5$ on the same gate, but at $R{=}5$ that interval reaches only 0.566, so neither clears the registered threshold on its own, and campaigns are not pooled.

No replicate at any scale triggered the one-sided-panel condition or needed the fallback, and a replicate took a median of 0.27, 9.25 and 63.09 hours from start to exit at p=500, 5000 and 10^4 (Table 9).

I.5 THE MISSING-DATA CHECK

The check reruns the design of Section 4.3 at $(p, n_{\text{total}}) = (500, 10^4)$ on ten further seeds, each once with complete data and once with missing entries, so the two views differ only in the missingness. The replicates follow the fixed stratum order: three ordinary (seeds 677092309, 1060767933, 206612481), three near-instrument (1958779806, 1207811676, 1934777877), two collider-descendant (197450974, 1689002239) and two strong exposure-descendant (80400408, 2003921685). Every variable except X is missing with probability $0.02 + 0.23\,\sigma(a + 0.7z)$, where z is the standardised exposure, σ the logistic function and a sets the overall rate to 10%, so missingness is at random given the fully observed exposure. The realised rates are 10.2% for Y, 10.0–10.1% for the mediators and 9.6–10.5% for the covariates.

Discovery completes its half-sample by K-MICE with K_{imp}=64, Bayesian ridge, posterior sampling and ten sweeps (Appendix G.1). The effect stage fits K-MICE with every other column as a predictor within each training fold, completing C first with posterior sampling and at most ten sweeps and then M and Y, and only transforms the held-out fold. In all ten replicates the compiled and supplied arms run under fold-wise completion and the unadjusted arm under observed-row weighting (Appendix C).

In both views every replicate reaches estimation with finite contrasts, recovers the target path and compiles a set that equals the supplied one and the minimal adjustment set of the true graph. Missingness raises L_{rm} by 2.33 percentage points, in all ten replicates, and leaves the unadjusted arm almost unchanged (Table 12). As in Table 3, the errors are those of the plug-in contrasts. A replicate takes about 9.3 minutes in the missing-data view.

Table 12: The missing-data check at p=500, compiled-set arm: the same ten seeds with complete data and with 10% missingness at random in every variable but the exposure. The interval is a two-sided 95% t interval over replicates.

Data	Target	Graph-adm.	MAE (ind./dir./total)	L_{rm} (95% interval)	MaxAE	Unadjusted L_{rm}
complete	10/10	10/10	.0096/.0107/.0170	5.08% (3.69–6.47)	.0333	44.52%
10% MAR	10/10	10/10	.0137/.0159/.0296	7.41% (5.83–8.99)	.0587	44.05%

I.6 STRESS TESTS OF THE BRIDGE

The campaigns of Table 3 give the compiler graph families that agree about adjustment near the target. The stress tests make the family disagree, both when a common valid set exists and when none does, and report every dataset of every experiment.

Designs with a common valid set. A shared core carries $X \to M$ (0.6), $M \to Y$ (0.5) and $X \to Y$ (0.2), with confounding $W_1 \to X$ (0.5), $W_1 \to W_2$ and $W_2 \to W_3$ (0.8 each) and $W_3 \to M$ (0.5), and a weak shortcut $W_1 \to M$ (0.10; 0.12 in the S2-reference stratum). The shortcut is detected in some graphs and not in others. Graphs that miss it also accept $\{W_2\}$ or $\{W_3\}$, so the family disagrees, while only $\{W_1\}$ is valid in every graph. The S2 and S2-reference strata add a second confounder U with $U \to X$ and $U \to Y$ (0.5 each), so the true adjustment set is $\{W_1, U\}$ there and $\{W_1\}$ in S1. Variable labels are permuted in each dataset, independent noise variables fill the table to p=50 or 500, n_{total}=10^4 is split evenly between discovery and effects, the data are complete, and the prior registers only the baseline pool W_1, W_2, W_3 and U, leaving the order of every other variable unknown. Experiments A (p=50) and B (p=500) use this linear core and a fixed path. Experiment C (p=500) makes the mediator–outcome channel $g(m) = c_{\text{lin}}m + \gamma\phi_k(m)$, with ϕ_k the linear, tanh, sine or radial-basis shape of the mechanism panel rescaled to M, η=0.95, c_{lin}=0.025 and γ set so that the indirect effect is 0.3. It adds a competing true path $X \to M_2 \to Y$ (0.5 and 0.3), registers M, M_2 and three decoys (a child of X, a parent of Y and a noise variable) in random order, and selects the path by the end-to-end rule of Section 3.3. The indirect, direct and total effects are 0.3, 0.35 and 0.65 on the main path and 0.15, 0.5 and 0.65 on the competing

one, and the true adjustment set is valid for both. Linear truths are analytic; nonlinear truths use 201-point Gauss–Hermite integration and agree with a 4×10^5-row Monte Carlo check to within one Monte Carlo standard error.

Designs without a common valid set. The boundary core has $X \to M$ (0.6), $M \to Y$ (0.5), $X \to Y$ (0.2) and a confounder Z with $Z \to X$ and $Z \to Y$ (0.5 each), so the indirect, direct and total effects are 0.3, 0.2 and 0.5 and the true adjustment set is $\{Z\}$. In N1 the public order of Z is unknown and $X \to Z$ lies in the same equivalence class: graphs with $Z \to X$ need Z, while in graphs with $X \to Z$ it descends from the exposure and the empty set is valid, so no set is valid in both. In the control N0, Z is registered as a baseline variable, which forbids $X \to Z$, and $\{Z\}$ is a common valid set. Boundary A and C use p=50 and boundary B uses p=500.

Method. Every stress test draws its family from fifteen row bootstraps of FGES under the degenerate-Gaussian BIC with penalty discount 4, each contributing two consistent extensions, for thirty graphs, followed by parent-set refinement of M and Y. Unlike the single-search family of Section 3.3, the bootstraps let the family disagree about weak edges, which is the condition these tests need. The compiler of Appendix B runs with K=2 and Q=12. A single-graph baseline reads the first graph alone and issues the first set in the same candidate domain that certifies on it. The effect stage of Section 3.5 runs with five outer folds for the compiled, single-graph, supplied and unadjusted arms, arms with the same set share one fit, and accuracy is that of the plug-in contrasts. A set is valid when it passes the joint admissibility check on the true graph.

Results. Where the family disagrees, the compiled set is valid in 147 of 148 datasets and withheld in the other, and none is invalid, which bounds the invalid rate by 2.0% at one-sided 95% confidence (Table 13). The single-graph baseline errs in 41 of the 180 datasets with a common set; the compiler corrects 35 of them and never errs where the baseline is right. Its accuracy matches that of the supplied true adjustment set (Table 14). In experiment C the main path is selected in all 120 datasets, the competing true path is supported by all thirty graphs in every dataset but ranks below it, and no decoy is ever supported; Table 15 breaks the experiment down by shape and stratum. Without a common set, no set is issued in any of the 60 N1 datasets, which bounds the invalid rate by 4.9% at one-sided 95% confidence, and all 40 controls issue the valid $\{Z\}$ (Table 16).

Table 13: Stress tests with a common valid set: sets issued and valid, withheld, or issued and invalid, over all datasets and over those where the thirty graphs disagree about adjustment. The last two columns compare with the single-graph baseline.

Experiment	n	Disagree	Disagree: valid / withheld / invalid	Valid	Invalid	Withheld	Single-graph errors	Corrected / reversed
A: p=50, linear	30	24	24 / 0 / 0	29	1	0	6	5 / 0
B: p=500, linear	30	22	22 / 0 / 0	29	1	0	10	9 / 0
C: p=500, nonlinear	120	102	101 / 1 / 0	115	4	1	25	21 / 0
Total	180	148	147 / 1 / 0	173	6	1	41	35 / 0

Table 14: Accuracy in the stress tests with a common valid set, compiled-set arm, with the columns of Table 3; values after ± are standard errors over datasets. The difference to the supplied true adjustment set is paired within dataset and given in percentage points. Experiment C excludes its one withheld dataset.

Experiment	L_{rm} (%)	Difference to supplied	MAE (ind./dir./total)	MaxAE	$\max_q \lvert\Delta\mathrm{MAE}_q\rvert$	Unadjusted L_{rm} (%)
A	4.09 ± 0.40	0.16 ± 0.16	.0098/.0099/.0126	.0362	.0006	42.75 ± 3.47
B	4.82 ± 0.62	0.09 ± 0.09	.0100/.0126/.0109	.0458	.0004	43.16 ± 3.23
C	3.95 ± 0.19	-0.00 ± 0.01	.0116/.0141/.0156	.0559	.0002	27.10 ± 0.88

Table 15: Experiment C by outcome shape and by stratum (120 datasets). Main path counts datasets whose selected path is the main one.

Group	n	Main path	Disagree	Valid	Invalid	Withheld	Single-graph errors	Corrected / reversed	L_{rm} (%)
linear	30	30	25	30	0	0	5	5 / 0	4.23 ± 0.43
tanh	30	30	26	28	2	0	6	4 / 0	3.65 ± 0.41
sine	30	30	25	30	0	0	6	6 / 0	3.76 ± 0.37
radial basis	30	30	26	27	2	1	8	6 / 0	4.14 ± 0.32
S1	40	40	38	39	1	0	11	10 / 0	3.99 ± 0.36
S2	40	40	35	37	3	0	11	8 / 0	3.94 ± 0.36
S2-reference	40	40	29	39	0	1	3	3 / 0	3.90 ± 0.28

Table 16: Stress tests without a common valid set (N1) and their controls with the common set $\{Z\}$ (N0). Must withhold counts N1 datasets whose family contains both orientations of the X–Z edge, so that no set certifies on every graph.

Experiment	N1 invalid	N1 withheld	N1 both orientations	N1 must withhold	N0 valid $\{Z\}$
A: p=50	0/10	10/10	8/10	8/10	10/10
B: p=500	0/10	10/10	10/10	10/10	10/10
C: p=50, further seeds	0/40	40/40	38/40	38/40	20/20
Total	0/60	60/60	56/60	56/60	40/40

Failure modes. Every invalid set comes from one source. In six datasets with a common set (one in A, one in B and four in C, two each for the tanh and radial-basis shapes) the weak shortcut $W_1 \to M$ is absent from all thirty graphs, so the family agrees on a set that is invalid in the true graph, and the compiler certifies it on every graph; the single-graph baseline errs in the same datasets. This is a detection limit of discovery for a weak edge, not a disagreement the compiler could resolve. One dataset of experiment C (radial basis, S2-reference) is withheld because one graph has no nonempty valid set of at most two variables, although the single-graph baseline happens to be right there. Four N1 datasets are withheld without being forced to: in two of boundary A the family consistently orients $X \to Z$ and a common set exists, and in two of boundary C a common set of at most two variables exists but lies outside the screened candidate domain; the single-graph baseline is right in three of these four. In N1 the single-graph baseline never issues an invalid set either, because it issues one only when the first graph admits a valid set.

Protocol. Every plan was fixed before its data were generated, on fresh random seeds excluded from all earlier plans and recorded with the plan; the servers only executed fixed plans, the truth was revealed at the audit after every dataset had finished, and denominators are fixed, so withheld and invalid sets stay in them. For experiment B and its boundary design, a pre-declared interim audit revealed the truth of the first replicate of each stratum before the others ran, and those replicates stay in the denominators. Two technical retries after server-environment failures reused the same seeds; the failed attempts had revealed the supplied and unadjusted effect truths of those first replicates, but the compiled arm had produced no output. The audit of experiment C ran the fixed audit function in parallel and matched its serial output on test data.

J THE HPP CAMPAIGN

This appendix gives the design and the full results of the Human Phenotype Project campaign of Section 5. Every count of participants is rounded down to the nearest ten.

J.1 THE REGISTERED PLAN

Table 17 is the plan in full. A route pairs one candidate mediator pool with one outcome, and the 64 routes are the 59 combinations of the original outcome-by-pool matrix together with five later additions from the blood and clinical-chemistry pool. The five pools are serum lipidomics (L), Nightingale NMR metabolites (N), preselected blood counts and clinical chemistry (B), fat

distribution and body composition (A), and blood-pressure and peripheral vascular measures (V). The exposure is the same in every route. Routes were run in six registered batches (Table 18), in an order fixed in advance and not changed by any result. Fifty-seven routes returned estimates, and together they cover all five pools, all nine domains and all twenty-four outcomes.

Table 17: The 64 registered routes. Each filled cell is one route, a mediator pool paired with an outcome, and gives the number of its candidate paths that returned all three effects; a blank cell is not a registered route. Pools are serum lipidomics (L), NMR metabolites (N), blood and clinical chemistry (B), body composition (A) and blood-pressure and vascular measures (V).

Domain	Outcome	L	N	B	A	V
Body composition	Total-body fat percentage	0	150			
	VAT volume	159	101			
Liver	Liver attenuation coefficient	0	120		25	
	Liver elasticity	501	59		80	
Blood pressure and vascular	Seated systolic blood pressure	537	0	6		
	Seated diastolic blood pressure	614	144			
	Carotid IMT, left	316	84			33
	Thigh-to-ankle PWV, right	208	102	0		
	Thigh-to-ankle PWV, left	150	158	0		
	ABI	1,018	108			
Sleep	Apnoea–hypopnoea index	96			0	
	Sleep efficiency	592			57	
Retina	Artery vessel density, left	286	86			22
	Artery distance tortuosity, left	153	94			11
Kidney	Serum creatinine	109	58			
	eGFR	172	64			
Muscle function	Hand-grip strength, right	1,389	168		85	
Bone	Femoral-neck BMD, left	354	107			
	Lumbar L1–L4 BMD	316	80			
Glycemic and metabolic	Laboratory glucose	84	77	0	28	
	HbA1c	221	66	7	49	
	CGM mean glucose	228	102		50	
	CGM coefficient of variation	305	72		49	
	New-onset diabetes	239	80		32	
Paths with estimates per pool		8,047	2,080	13	455	66

J.2 Cohort, timing and samples

The exposure is taken at the baseline visit of the HPP. For each route the mediator record is the one closest to the stool sample within 45 days; lipidomics, NMR, body composition, blood pressure and vascular measures are restricted to the baseline visit, and blood tests, which carry no visit label, are matched by date alone. The baseline outcome is the record within 30 days of the mediator sample, and the follow-up outcome is the record 180 to 1,095 days after it that lies closest to 730 days, the earlier one on a tie. Where one time point holds several conflicting records, the participant is unavailable for that route rather than one record being chosen.

Participants are split into a discovery half and an effect half by a fixed hash of the participant and cohort identifiers, nominally $50{:}50$, and every batch uses the same split; 4,310 participants of the discovery half and 4,370 of the effect half have a baseline metagenome. A path uses, in its effect half, the participants with the genus, the mediator and both the baseline and the follow-up outcome; its discovery half does not require the follow-up outcome, except in the first batch. Covariates need not be complete, and mediators are taken one at a time, so a participant need not be complete on the whole pool. Table 18 gives the resulting sample sizes.

J.3 VARIABLES

Exposure. The gut microbiome is profiled by shotgun metagenomics and summarised at genus level with MetaPhlAn 4. Each genus enters as its pivot log-ratio balance against the other genera of a fixed reference set of 908 genera, which equals a rescaled centred log-ratio coordinate. Zeros are replaced by one constant, half the smallest positive relative abundance in the discovery reference matrix (4.5×10^{-5} per cent), which is fixed on the discovery half and not re-estimated by route. In all, 291 genera appear as candidate exposures.

Mediators. The lipidomics pool registers 187 lipids, of which 163 to 183 are available on a route (99 on the laboratory-glucose route); the NMR pool registers 250 measures, 249 of them available; the blood pool registers 40 measures, 39 on the glucose route, which excludes glucose itself as a mediator; the body-composition pool holds total-body fat percentage and VAT volume; and the blood-pressure and vascular pool holds seated systolic and diastolic blood pressure and right and left thigh-to-ankle PWV, two of them on each route.

Outcomes. Every outcome is a released field except two: eGFR, derived from serum creatinine, and new-onset diabetes. Carotid IMT is the left-side measure; the ABI uses the released field `r_abi`, whose documentation conflicts on its side, so no laterality is claimed; the retinal fields are named for the left eye, but the side has not been verified; bone mineral density is measured at the left femoral neck and at L1–L4; grip strength is that of the right hand; liver elasticity is the released aggregate over quality boxes; and the sleep outcomes are single-night records. New-onset diabetes is an event among participants free of diabetes at baseline, that is, with HbA1c below 6.5% (48 mmol/mol); the event is an HbA1c of at least 6.5% at follow-up, and its effects are differences in risk.

Covariates. The 35 baseline covariates are approximate age, sex, the baseline outcome, season as the sine and cosine of the mediator sampling month, 20 genetic principal components, BMI from the latest dated anthropometry and nine lifestyle variables from the latest dated questionnaire: smoking in four levels, alcohol frequency, weekly minutes of walking, of moderate and of vigorous activity, processed meat, oily fish, added salt and a vegetarian diet. Categorical covariates enter as $k - 1$ contrasts, and the single-level cohort adds no column. Every model adjusts for all 35 and the path's compiled adjustment set; for body-composition mediators this makes the effect conditional on BMI.

Missingness by block. Each route has its own analysis population, so missingness is reported by block rather than as one overall rate. The exposure is never missing; zeros, about 88% of the reference matrix, are not treated as missing and are replaced as above. Within its discovery pairing, a mediator field is missing for a median of 11.6% of participants in the lipidomics pool (range 0–98.4%), 0% in the NMR pool (0–87.7%) and about 24% in the blood pool (13.1–66.2%), for 1.4% and 2.2% in the body-composition pool, and for 0% (blood pressure) and 6.2–6.4% (PWV) in the vascular pool; a path keeps only participants whose mediator is observed. Outcomes are likewise complete within each analysis population, since a path requires both the baseline and the follow-up outcome; the follow-up is the binding requirement, which is why the effect half keeps fewer participants than the discovery half (Table 18). Only covariates are imputed: age, sex, the baseline outcome and season are complete, BMI is missing for under 1% of participants and the genetic principal components for at most about a fifth, and missingness concentrates in the questionnaire, where vigorous-activity frequency is missing for at least half of the participants, added salt for about a fifth to three-fifths and every other lifestyle field for at most about a third.

Covariate completion. Missing covariates are completed by K-MICE with every other covariate as a predictor, separately in the two stages and never across them. Discovery completion is fitted within each discovery half-sample over five sweeps, with Bayesian ridge for numerical and logistic regression for categorical targets, unlike the Gaussian updates of Appendix G.1, and draws from their predictive distributions, under seed 110990908 from which the per-mediator and per-half-sample seeds are derived. Effect completion is fitted on each training fold of the effect half and only applied to its validation fold, over five sweeps with Bayesian ridge, no posterior sampling and seed 909202641 plus the fold index. Both stages use five sweeps rather than the default ten (Appendix G.1), and no missing-data mode label is assigned on this campaign.

J.4 Discovery and candidate paths

Within each route, discovery runs on the discovery half exactly as in Sections 3.2–3.3, with the screen, search and family settings of the end-to-end runs (Appendix G.3): the completed table is screened to each variable's 32 strongest correlation partners together with the registered role pairs, FGES searches the screened domain under the registered role order genus $\prec$ mediator $\prec$ outcome, and $S{=}30$ consistent extensions are refined into the family. Every genus–mediator–outcome path with $\widehat{\mathrm{PSS}} \geq \tau = 0.2$ becomes a candidate path of the route. For every candidate path the compiler of Section 3.4 issues an adjustment set of at most two variables ($Q{=}12$, $K{=}2$), and every model adjusts for it together with the 35 baseline covariates. Unlike the end-to-end runs, which carry only the highest-ranked path forward, every candidate path is estimated.

J.5 Estimation, guards and multiple testing

The shift is one standard deviation of the exposure: the exposure is the standardised genus balance $X = (B - \hat{\mu}_D)/\hat{\sigma}_D$, which the shift sets to $X + 1$, with $\hat{\mu}_D$ and $\hat{\sigma}_D$ fixed on the discovery reference sample of each route and reused unchanged on its effect half, so the shift on the balance scale is $\hat{\sigma}_D$ and may differ between routes. The outcome regression is the channel-additive class of Section 3.5. The exposure densities given C and given (M, C) are homoskedastic Gaussian working models, with network conditional means and a variance set by the residual root mean square on a calibration split inside each outer training fold, fitted separately for every route, fold and model; the Gaussian form is a working assumption.

The three effects are the population intervention direct, indirect and total effects of this shift (Díaz & Hejazi, 2020), the stochastic-shift counterparts of NDE, NIE and TE, and the direct and indirect effects add to the total. Unlike the natural effects of Section 2, they need no cross-world condition: (A1) to (A3) of Appendix A, with positivity requiring the shifted exposure to stay within the support of the observed one, identify them (Díaz & Hejazi, 2020). For each path the direct, indirect and total effects are estimated by the cross-fitted one-step estimator, with the standard error taken from the empirical variance of its efficient-influence-function scores, Wald intervals $\hat{\theta} \pm 1.96\,\widehat{\mathrm{SE}}$ and two-sided normal p-values. These are asymptotic and rest on the nuisance-rate and positivity conditions of the one-step estimator, including the Gaussian working densities used for the exposure. For these shift effects the one-step estimator replaces the plug-in contrasts of Section 3.6, and its intervals are distinct from the row-score variance there, which carries no coverage claim.

Guards. A path is estimated only when its samples meet the registered minimums: at least 200 participants in discovery, 125, 150 or 200 in the effect half according to the route tier, and at least 100 training and 25 test rows in every cross-fitting fold. It is then refused rather than estimated when its weights fail the estimation guards: an effective sample size of the outcome weights or of the shift ratio below a floor, a largest weight above a catastrophic cap, or too heavy a tail beyond the routine cap. Of the 7,219 refused paths, 94.0% fail one of these weight guards, 5.4% the minimum sample or row alignment, 0.5% an exposure scale that cannot be estimated in discovery and 0.1% a density ratio that is not finite; a refusal can carry several reasons.

Multiple testing. Each batch is one family, whose size K_{fam} is its number of candidate paths, including refused paths and the 132 paths across batches that were neither estimated nor refused. Benjamini–Hochberg is applied to the indirect-effect p-values with $m = K_{\mathrm{fam}}$, and Holm's procedure to the $3K_{\mathrm{fam}}$ p-values of the three effects. All adjusted values are computed on the final table. Table 18 gives the counts; nominal $p < 0.05$ holds for 735 of the 10,661 indirect effects (6.9%).

Table 18: The registered batches. Routes counts the routes with at least one estimated path, K_{fam} is the number of candidate paths, which is the size of the batch family, and Estimated and Refused count paths. Sample sizes are ranges over the estimated paths, rounded down to tens. BH counts the indirect effects that pass Benjamini–Hochberg at 0.05 within the batch.

Batch	Routes	K_{fam}	Estimated	Refused	Discovery N	Effect N	BH
P1	1	948	614	334	470–1,050	500–1,050	2
P2	4	1,774	1,028	743	470–1,260	130–270	2
P3	8	6,472	3,977	2,385	370–2,560	120–1,040	3
P4A	3	789	360	427	510–660	150–240	1
P4B	3	347	198	149	2,200–2,480	170–220	3
P5	38	7,682	4,484	3,181	460–2,400	160–970	1
Total	57	18,012	10,661	7,219	370–2,560	120–1,050	12

J.6 The significant paths

Table 19 lists the 12 indirect effects that pass Benjamini–Hochberg at 0.05 within their batch, grouped by system. Each runs from a genus through a serum metabolite or a fat depot to an outcome.

Table 19: The 12 indirect effects that pass Benjamini–Hochberg at 0.05 within their batch, grouped by system. The indirect effect is the change in the mean follow-up outcome under a one-standard-deviation shift of the genus balance, in the unit of the outcome, with its nominal Wald interval. BH is the adjusted p-value, † marks the one path that also passes Holm across the three effects, and N is the effect sample, rounded down to tens.

Genus	Mediator	Outcome	Batch	Indirect effect [95% CI]	BH	N
Metabolism, muscle and sleep						
Prevotella	Total BCAA	Glucose (mg/dl)	P5	3.32 [1.93, 4.71]	0.023	180
GGB21699	TAG 58:8	Grip strength (kg)	P3	1.93 [1.10, 2.75]	0.030	420
GGB81451	TAG 51:0	Sleep efficiency (%)	P3	−1.72 [−2.52, −0.930]	0.045	180
Liver and visceral fat						
unclass. *Christensenellaceae*	VAT volume	Liver elasticity (kPa)	P4B	−0.529 [−0.768, −0.291]	0.005†	170
Oscillibacter	VAT volume	Liver elasticity (kPa)	P4B	−0.629 [−0.940, −0.318]	0.013	170
unclass. *Christensenellaceae*	VAT volume	Grip strength (kg)	P4B	1.74 [0.788, 2.70]	0.040	220
Blood pressure and vessels						
Alistipes	TAG 54:4	Diastolic BP (mmHg)	P1	−2.01 [−2.93, −1.09]	0.018	660
Alistipes	DAG 36:2	Carotid IMT (mm)	P2	−0.0522 [−0.0764, −0.0279]	0.022	160
Alistipes	TAG 54:4	Carotid IMT (mm)	P2	−0.0505 [−0.0734, −0.0275]	0.022	180
Akkermansia	VLDL-TG	ABI (ratio)	P4A	0.0429 [0.0222, 0.0635]	0.037	180
unclass. *Peptostreptococcaceae*	AcCa C16:0	ABI (ratio)	P3	−0.0116 [−0.0168, −0.00643]	0.037	640
Alistipes	DAG 36:2	Diastolic BP (mmHg)	P1	−1.09 [−1.63, −0.541]	0.045	940

J.7 Why the causal status is withheld on this cohort

The causal status of a record is set only when the estimand is identified on a graph that is accepted as causal, and two things stand in the way here. First, no ground-truth graph exists for this cohort, so no structure recovered on it is certified as causal; the refined family is an ensemble of candidates, and nothing in this paper claims any member of it to be the causal graph. Second, what the compiler certifies are the single-world back-door statements of equation 3, which support (A3) on a graph already accepted as causal but do not reach that acceptance. The field is therefore left unset on all 64 routes rather than filled in optimistically, and the fields that are filled in carry what a run does establish: that the estimator executed out of fold against a fixed set, and that the set certified on the path-supporting members of the family. This is the convention of Appendix A.

K Limitations

Identification. Causal interpretation of any estimate requires identifying assumptions beyond graph checks. The compiler certifies single-world back-door statements on the sampled graphs;

it establishes neither the cross-world condition nor that any graph in the family is causal (Appendices A and B).

Scope of the simulations. End-to-end evidence is simulated, with complete public tier orders and at most two adjustment variables per path, and an edge missed by every sampled graph cannot be repaired by the compiler (Appendix I.6).

Uncertainty. The effect variance of Appendix D is a diagnostic without a coverage guarantee.

Real data. The HPP campaign has no ground-truth graph. Its effects are stochastic-shift effects with nominal Wald intervals, multiplicity is controlled within registered batches, and the chains of Section 5 are candidates that agree with reported links rather than confirmed mechanisms (Appendix J.7).