TOPOLOGY-CONDITIONED FEASIBLE-SET GENERATION FOR CONTINUOUS-TIME MULTIGRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating temporal subgraphs requires changing interaction structure while preserving observations and quantitative constraints. Structural edits alter relay equations and can make pinned observations incompatible with positive flow. We formulate this as *topology-conditioned feasible-set generation*. Each structural proposal induces its own constraint system, and the minimum distortion required to complete it shapes which structure is selected. A typed edit grammar and a policy trained to reconstruct corrupted skeletons propose and score admissible changes while protecting pinned events. A conditional diffusion prior then generates quantities and event times for each candidate topology. Each candidate is completed by **Dirichlet-Boundary-Anchored KL projection** (**DBA-KL**), which finds the closest positive flow that preserves pinned quantities and satisfies relay conservation. Under strict feasibility this completion is unique, and its KL distortion separates into the minimum change the constraints require plus any extra distortion from another feasible completion. All completed outputs preserve positive quantities and pinned observations, and every output containing a relay satisfies the relay-balance tolerance. On three anti-money-laundering benchmarks from a single simulator family, our DBA-KL configurations attain higher four-classifier average precision than the ECI-DBA and PCFM-log adaptations on all three. Relative gains reach 23.5% over ECI-DBA and 12.7% over PCFM-log on LI-Medium, and the terminal variant exceeds real-data training on HI-Small. Downstream accuracy alone is nevertheless a poor guide to constrained generation: an ablation that removes flow completion scores higher on every dataset while violating relay balance on 610 of 610 audited relay-containing outputs. Predictive utility and quantitative validity therefore have to be measured separately.

1 INTRODUCTION

Continuous-time interaction multigraphs record who interacts with whom, when, and with what quantity; transactions, communication events, and power flows are typical instances (Altman et al., 2023; Marot et al., 2021). Predictors on such data are increasingly trained and evaluated with generated interactions, from augmentation of dynamic graph models (Tian et al., 2024) and simulated anti-money-laundering benchmarks (Altman et al., 2023) to network planning, distribution-shift benchmarks, and counterfactual explanation (Kan et al., 2025; Gui et al., 2022; Lucic et al., 2022). Graph generative models increasingly supply those interactions (Xu & Ma, 2025). Many of these uses need scenarios that differ from the observed data in their *relationships*, not only in their attribute values, because the decisions being trained or tested concern which entities interact.

Our setting narrows this to a source subgraph whose observed evidence must be kept. A practitioner-reported financial-risk bottleneck, which includes authorized red/blue-team testing of LLM-assisted analysis (a qualitative report, not a production measurement; Appendix F), illustrates the problem: varying amounts and timestamps on familiar transaction skeletons leaves decisions about different accounts and subgraphs untested, while the observations defining a case must stay fixed. Attribute-only generation and identifier renaming leave incidence unchanged, so structural diversity (Velikonivtsev et al., 2024) has to be produced deliberately and assessed separately from representation effects (Fatemi et al., 2024). We study this as *structural scenario generation*: local, quantitatively valid variations in connectivity, amounts, and event times that preserve specified observations.

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.

Structural edits threaten quantitative validity. Under our lossless model with zero net accumulation, flow must balance at every relay of the modeled skeleton. Inserting, deleting, or rewiring an event therefore alters the balance equations, and the fixed observations can become incompatible with any positive balanced flow. Even when a valid completion exists, the proposed amounts must be adjusted to reach it. A validity check alone cannot separate the adjustment the constraints require from the excess distortion of the chosen completion, which matters when scenarios are compared, audited, or selected.

Existing methods and why completion must be accounted for. Joint structure–attribute generators (Jo et al., 2022; 2024; Zhou et al., 2024; Vignac et al., 2023; Qin et al., 2025) and temporal interaction generators (Zhou et al., 2020; Hosseini et al., 2025) sample whole graphs from a learned distribution without fixing observed events or evaluating topology-induced flow constraints. Constraint-aware generators (Madeira et al., 2024; Hansen et al., 2023; Christopher et al., 2024; Zhang et al., 2026; Shankar et al., 2026) enforce a constraint set fixed before sampling. Here the constraint set is itself the outcome of a structural choice: each candidate topology induces its own balance system, its own feasible set, and its own minimum adjustment from the same frozen proposal. Feasibility and distortion are therefore computed per candidate, and they shape which structure is emitted.

ECI (Cheng et al., 2025) and PCFM (Utkarsh et al., 2025) are the closest alternatives to our completion step. Both can be adapted to enforce relay conservation, but their objectives differ from ours. ECI repeatedly changes its correction reference during sampling, which makes the final distortion difficult to attribute to a single proposal, and every round costs another flow solve. PCFM can use a fixed terminal reference, but enforcing positivity in log space leads to a nonconvex projection that is certified only to stationarity. Neither separates the adjustment a structure requires from the excess of a particular completion. DBA-KL instead uses one fixed KL reference and returns the minimum-distortion positive feasible completion. The KL projection itself is classical (Csiszár, 1975; Bregman, 1967); our contribution is how it is used for topology-dependent generation. Its analytic relay updates preserve positivity without a Jacobian solve, and the objective is convex with a unique optimum under strict feasibility (Section 3.3, Appendix B). One solve per candidate against one fixed reference suffices. The total distortion of any positive feasible completion then splits exactly, with no cross term, into the minimum the constraints require plus that completion's excess. This property puts every candidate on one scale.

Framework, evidence, and contributions. A typed edit grammar that protects fixed events and a policy trained to reconstruct corrupted skeletons propose candidate topologies. A boundary-conditioned EDM prior (Karras et al., 2022) generates quantities and event times for each one. A single terminal DBA-KL solve and a temporal completion establish validity and cost, and selection weighs learned structural preference against completion cost, so feasibility and distortion shape the emitted structure (Section 3). We make three contributions. First, we formulate structural generation as selection among topology-dependent feasible sets: changing the topology also changes the continuous constraints. Second, we introduce a signed-incidence solver that finds the minimum-KL positive completion for each feasible topology. This gives every candidate a directly comparable completion cost. Third, we evaluate predictive utility and quantitative validity separately, and we show that high downstream accuracy does not guarantee valid generated scenarios. On one simulator-generated anti-money-laundering benchmark family (LI-Small, HI-Small, LI-Medium), every DBA-KL configuration attains higher four-classifier average AP than the ECI-DBA and PCFM-log adaptations. An ablation with no flow completion nevertheless scores higher still, while failing relay balance on every audited relay-containing output. Our empirical conclusions are limited to this benchmark family; the framework itself is stated for continuous-time multigraphs generally (Section 4).

2 PRELIMINARIES

2.1 CONTINUOUS-TIME INTERACTION MULTIGRAPHS

A continuous-time interaction multigraph is an event collection $\mathcal{G} = (V, \mathcal{E})$ whose events $e = (u_e, w_e, \tau_e, v_e, c_e)$ record a source entity, a destination entity, a real-valued occurrence time, a

strictly positive flow quantity, and observed categorical metadata; distinct events may share endpoints, so event multiplicity is retained rather than collapsed into a binary adjacency matrix. Quantities that enter one balance equation represent the same modeled quantity in common units, with a positive conversion table frozen before generation when native units differ (Appendix C.1). Given a source attributed skeleton $S_0 = (V_0, E_0, X_0)$ and observed context $\mathcal{H}_0$, we generate structural and attribute variants that preserve specified observations: a variant may insert, delete, or rewire unpinned events, change which observed entities participate, and generate continuous quantities and event times.

2.2 Observation boundary and skeleton-local conservation

The observation boundary $\mathcal{B}$ comprises the frozen context $\mathcal{H}_0$, the designation of modeled and context-only events, the pinned events with their observed attributes, and the source time window. The modeled event set of a skeleton S is $E_S = E_S^{\mathrm{pat}} \dot\cup E^{\mathrm{att}}$: pattern events may be edited subject to the pins, fixed attachment events belong to the balance model but are never edited, and context-only events condition generation without entering any balance equation. The pinned pattern events $P \subseteq E_0^{\mathrm{pat}}$ keep their endpoints, quantities, timestamps, and prescribed categorical attributes in every admissible candidate; with the attachments they form the fixed event set $P_c = P \cup E^{\mathrm{att}}$, and $F_S = E_S^{\mathrm{pat}} \setminus P$ is the free event set, so $E_S = P_c \dot\cup F_S$. Boundary preservation means unchanged context observations and unchanged observations on P_c, which act as Dirichlet-type conditions on the event coordinates; unpinned events, including retained original events, remain editable subject to the structural grammar (pin selection: Appendix C.1).

Relays are the entities with positive in- and out-degree in the candidate's own modeled incidence, $V_R(S) = \{a : \deg^-_{E_S}(a) > 0 \text{ and } \deg^+_{E_S}(a) > 0\}$, and under our lossless, zero-net-accumulation model the incoming and outgoing quantities balance at each relay, $\sum_{e \in E_S:\, w_e = a} v_e = \sum_{e \in E_S:\, u_e = a} v_e$. With the relay-restricted signed incidence $[A_R(S)]_{a,e} = \mathbf{1}\{w_e = a\} - \mathbf{1}\{u_e = a\}$, eliminating the fixed quantities $\bar v_{P_c}$ leaves the linear system $A_{R,F_S}(S)\, v_{F_S} = b_S$ with $b_S = -A_{R,P_c}(S)\, \bar v_{P_c}$, whose right-hand side depends only on fixed modeled quantities. Feasibility is not automatic: inconsistent fixed observations or a structural edit can leave no strictly positive solution. Each candidate structure thus induces its own feasible set of positive, balanced quantities and admissible times, and structural admissibility and continuous feasibility are separate conditions; Appendix A formalizes this topology-dependent domain, the conventions for parallel, self-interaction, and relay-free cases, and the notion of structural coverage that the evaluation uses.

3 Methodology

Figure 1 summarizes the framework. Given a source skeleton S_0 and observation boundary $\mathcal{B}$, a typed edit grammar and a learned structural policy construct and score a bounded menu of admissible candidates that preserve the fixed events (Section 3.1); a boundary-conditioned diffusion prior proposes quantities and event times for each candidate, with the topology fixed within each denoising trajectory (Section 3.2). DBA-KL then completes each proposal to the positive, relay-balanced assignment of minimum KL distortion, which establishes the candidate's validity and its minimum required distortion (Section 3.3). Finally, temporal completion and selection combine learned structural preference with the completion costs of the accepted candidates, so feasibility and distortion influence which topology is emitted. Every component is fitted and then frozen (Section 3.4). Throughout, a candidate e is a structural edit of S_0 with skeleton S_e; all modeled observations on $P_c = P \cup E^{\mathrm{att}}$ are preserved and only the free events $F_e = E_{S_e}^{\mathrm{pat}} \setminus P$ are optimized, while context-only events condition the prior but do not enter relay equations. Completion is applied once per candidate, after denoising; Algorithm 1 (Appendix C.1) summarizes one generation, and Appendix G places the design among related methods.

3.1 Admissible structural generation

Problem. Most edits of a source skeleton alter fixed events, attach identities never observed near the source, or duplicate one another, so the generator needs a bounded admissible menu and a quantity-free preference over it before any quantities or times exist.

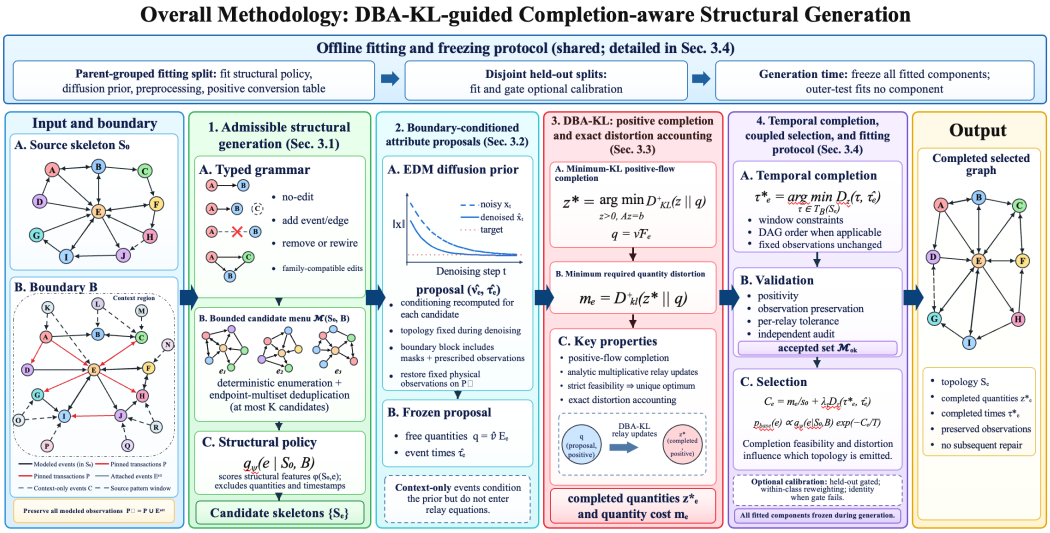

Figure 1: Overview of the framework. Top: the structural policy, diffusion prior, preprocessing, and conversion table are fitted on a parent-grouped split, the optional calibration is fitted and gated on disjoint held-out splits, and every component is frozen at generation time. Bottom, for a source skeleton S_0 with boundary $\mathcal{B}$ (pinned transactions, attached events, context): (1) a typed edit grammar enumerates a bounded candidate menu scored by the structural policy $q_\psi(e \mid S_0, \mathcal{B})$; (2) a boundary-conditioned EDM prior proposes quantities and times for each candidate with its topology fixed during denoising; (3) DBA-KL completes each proposal to the unique positive relay-balanced optimum $z_e^\star$ and records the minimum required distortion m_e; (4) temporal completion, validation, and selection with cost $C_e = m_e/s_0 + \lambda_\tau D_\tau$ emit one completed graph with all modeled observations preserved and no subsequent repair.

A typed grammar constructs a bounded menu $\mathcal{M}(S_0, \mathcal{B})$ containing the no-edit state and family-compatible changes to unpinned pattern events. Edits preserve P_c and use identities observed in the source or its context. Deterministic enumeration and endpoint-multiset deduplication retain at most K candidates; this menu is not the full feasible structural support. A log-linear policy scores standardized structural features $\phi(S_0, e)$, excluding quantities and timestamps:

$$q_\psi(e \mid S_0, \mathcal{B}) = \frac{\exp\{\psi^\top \phi(S_0, e)\}}{\sum_{e' \in \mathcal{M}(S_0, \mathcal{B})} \exp\{\psi^\top \phi(S_0, e')\}}. \tag{1}$$

The policy is fitted by reconstructing original topologies from synthetically corrupted skeletons under a parent-balanced negative log-likelihood with ridge regularization (Appendix C.2, Eq. (8)); it learns conditional reconstruction, not an observed temporal mutation law, and missing reconstruction targets are recorded as menu-coverage failures.

3.2 Boundary-conditioned attribute proposals

Problem. Each candidate needs quantities and times that are plausible and informed by the fixed observations. A neural prior cannot satisfy the relay-balance and ordering constraints exactly, so conditioning shapes the proposal and completion enforces the constraints.

Each event is encoded by a standardized log quantity and a scaled logit of normalized time. A shared row-wise MLP with EDM preconditioning (Karras et al., 2022) predicts clean coordinates from structural and categorical features, context summaries, and a boundary block of fixed-coordinate masks and prescribed observations. For each candidate, Euler sampling with keyed noise over a decreasing schedule overwrites the fixed coordinates on P_c with their observations at every step, restores them after decoding, and freezes the resulting proposal $(\widehat{v}_e, \widehat{\tau}_e)$ for completion and selection. Architecture, training objective, masks, schedules, encoders, and the sampling equation are in Appendix C.3.

3.3 DBA-KL: POSITIVE COMPLETION AND EXACT DISTORTION ACCOUNTING

Problem. A proposal generally violates the relay equations of its candidate skeleton, and an edit may leave no positive assignment at all; completion must return the closest positive balanced assignment whenever one exists and quantify how much of the change the constraints required, so that candidates and completion rules share one reference.

For candidate S_e, eliminate the fixed modeled quantities as in Section 2.2 and write $A = A_{R,F_e}(S_e)$, $b = b_e$, and $q = \widehat{v}_{F_e} > 0$, holding the skeleton, boundary, units, and proposal fixed. DBA-KL computes

$$z^\star = v^\star_{F_e} = \operatorname*{arg\,min}_{z>0,\ Az=b} D^+_{\mathrm{KL}}(z\|q), \qquad D^+_{\mathrm{KL}}(z\|q) = \sum_i \left[z_i \log \frac{z_i}{q_i} - z_i + q_i \right]. \tag{2}$$

This is an information projection of quantities, without probability normalization (Csiszár, 1975; Bregman, 1967). If some $z^+ > 0$ satisfies $Az^+ = b$, a unique positive global minimizer exists, including when relay equations are redundant. It supplies a reference for measuring completion distortion.

Exact distortion accounting. Every alternative $u > 0$ satisfying the same $Au = b$ obeys

$$\underbrace{D^+_{\mathrm{KL}}(u\|q)}_{\text{total distortion}} = \underbrace{D^+_{\mathrm{KL}}(z^\star\|q)}_{\substack{\text{minimum required} \\ \text{KL distortion}}} + \underbrace{D^+_{\mathrm{KL}}(u\|z^\star)}_{\text{excess KL distortion}}. \tag{3}$$

Define $m_e = D^+_{\mathrm{KL}}(z^\star\|q)$ and $\eta_e(u) = D^+_{\mathrm{KL}}(u\|z^\star)$. The first is the minimum adjustment needed to reconcile this proposal with its skeleton and fixed observations; the second is the additional KL cost of the chosen feasible allocation, so equal total distortions can reflect either proposal–constraint incompatibility or departure from the minimum-cost reference. The cross term vanishes by the optimality condition $\log(z^\star/q) = -A^\top \lambda^\star$ (Appendix C.5; numerical verification in Appendix E). This identity is a property of the projection and also holds for other completion rules on the same (A, b, q); it is not a task-loss decomposition or an identity for temporal inequalities.

Analytic completion. Starting at $z^0 = q$ with dual variable $\lambda^0 = 0$, DBA-KL cyclically maximizes the dual of Eq. (2) one relay row at a time. For relay row a, with I and O the current free totals over positive and negative coefficients, the exact coordinate update is

$$Ie^{-\delta} - Oe^{\delta} = b_a, \qquad z_i \leftarrow z_i e^{-\delta A_{a,i}}, \qquad \lambda_a \leftarrow \lambda_a + \delta, \tag{4}$$

where e^δ is the positive root of $Ot^2 + b_a t - I = 0$, and $\delta = \frac{1}{2}\log(I/O)$ when $b_a = 0$. Each update preserves positivity and maximizes one dual coordinate without a global Jacobian solve. Under strict feasibility, ideal cyclic updates converge to $z^\star$ (Proposition 4), since each exact row update satisfies a row-wise Pythagorean identity against any fixed positive feasible point (Proposition 3). Numerical acceptance requires positivity, stationarity, and every per-relay balance tolerance, with bounded-budget failures recorded as solver outcomes; stable root formulas, one-sided cases, and sweep cost are in Appendices C.4 and C.10.

3.4 TEMPORAL COMPLETION, COUPLED SELECTION, AND FITTING PROTOCOL

Problem. Event times must satisfy the window and ordering restrictions. One completed graph must be chosen so that topology and attributes are selected jointly. Every fitted component must be trained without access to evaluation cases.

Temporal completion, independent of the quantity completion, minimizes a window-normalized squared distance $D_\tau(\tau, \widehat{\tau}_e)$ between the free times and their proposals over $\mathcal{T}_{\mathcal{B}}(S_e)$, the set of time assignments meeting the window, pin, and ordering restrictions (Appendix A); the problem is convex, its minimizer is $\tau^\star_e$, and fixed observations remain unchanged (Appendix C.6). With the source scale $s_0 = \max(\sum_{i \in E_0^{\mathrm{pat}}} v_i^0, 10^{-9})$, each accepted candidate has cost $C_e = m_e/s_0 + \lambda_\tau D_\tau(\tau^\star_e, \widehat{\tau}_e)$; the scale normalizes distortion, not throughput, and time constraints apply even when $\lambda_\tau = 0$. Given frozen proposals $\widehat{\mathcal{X}}$ and the numerically accepted candidate set $\mathcal{M}_{\mathrm{ok}}$, selection draws

$$p_{\mathrm{base}}(e \mid S_0, \mathcal{B}, \widehat{\mathcal{X}}) = \frac{q_\psi(e \mid S_0, \mathcal{B}) \exp(-C_e/T)}{\sum_{e' \in \mathcal{M}_{\mathrm{ok}}} q_\psi(e' \mid S_0, \mathcal{B}) \exp(-C_{e'}/T)}, \quad e \in \mathcal{M}_{\mathrm{ok}}. \tag{5}$$

This profiles a realized finite menu rather than the full domain of Appendix A. It uses accepted numerical costs without subtracting an unknown excess term (Appendix C.5), treats an empty accepted menu as an omission, and applies no subsequent repair (Algorithm 1).

Fitting protocol. The prior, structural policy, preprocessing, and the optional calibrator with its gate are fitted on disjoint parent-grouped partitions of the outer-training data, never on outer-test cases, and are frozen during generation (Appendix C.1).

Calibration and comparison schedules. An optional held-out-gated calibrator reweights completed candidates within protected classes and is the identity when its gate fails (Appendix C.7). ECI-DBA (Cheng et al., 2025) (the same inner solver driven by changing proposals) and PCFM-log (Utkarsh et al., 2025) (a fixed-reference log-distance objective) are comparison adaptations, not steps of the generator (Appendix C.9).

4 EXPERIMENTS

We organize the evaluation around four questions: downstream predictive utility (RQ1), whether utility separates valid from invalid generation (RQ2), generation validity and structural coverage (RQ3), and computational cost and scaling (RQ4). Appendix E verifies the distortion identity of Section 3.3 numerically.

4.1 GENERAL EXPERIMENTAL SETTINGS

Benchmarks. Every reported experiment uses one simulator-generated benchmark. LI-Small, HI-Small, and LI-Medium are subsets of the same IBM-AML family (Altman et al., 2023), not independent application domains, and no production data is involved. The framework is stated for continuous-time interaction multigraphs generally; only the AML family is evaluated here, and transfer to another domain remains open (Appendix H).

Datasets, task, and evaluators. Accounts are nodes and transactions carry directed endpoints, quantities, timestamps, and categorical metadata. Supplied case lists define positive source skeletons, split temporally within each family with siblings kept together, and the generator's fitting, calibration, and gating partitions stay inside the outer training split (Section 3.4). The task is retrospective synthetic-positive replacement. Real/TRTR uses original training positives, a generated arm replaces them with its generated positives, and both retain the original normal background and the chronological 80/20 normal-target split. Evaluation uses the same intended original test graph and targets. Later-in-window context is permitted and test labels are excluded from fitting; case construction is annotation-informed. Appendix D.2 gives the exact graph and target sets and the adapter rules. The evaluators are EdgeSAGE, EdgeGCN, EdgeGAT, and EdgeGIN, scored on saved final-epoch test metrics, where AP is the native `PR-AUC` field. For seed s and configuration c the four-GNN average is $a_{c,s} = \frac{1}{4}\sum_m \mathrm{AP}_{m,c,s}$, and we report its mean and population SD across seeds. The grids contain 220 complete fits on LI-Small and HI-Small (five seeds per cell) and 176 on LI-Medium (four), with no missing fit.

Comparisons. Table 1 compares Real with four DBA-KL configurations (terminal Learned, Calibrated, fixed-topology, and three-round DBA-KL with projected EDM), ECI-DBA with one and three correction rounds, PCFM-log, and the equality-only controls ECI-naive and PCFM-naive (Appendix C.9). It also lists two external generators evaluated on their raw outputs, HARPOON (Shankar et al., 2026) and GDSS (Jo et al., 2022), which enforce neither relay balance nor pinned observations.

4.2 RQ1: DOES GENERATED DATA SUPPORT DOWNSTREAM PREDICTION?

Setting. Table 1 reports AP for EdgeSAGE and EdgeGIN; Table 5 (Appendix D.4) gives the four-GNN average and its SD. We select no configuration per evaluator.

Table 1: Downstream AP (native PR-AUC) under synthetic-positive replacement for the two strongest evaluators on every dataset, EdgeSAGE and EdgeGIN: means over detector seeds (five on LI-Small and HI-Small, four on LI-Medium), rounded to three decimals. Real is the original-data reference; dotted rules separate it, the DBA-KL configurations, the comparison adaptations, and the two external generators. Bold marks the best topology-changing DBA-KL variant in each column; ties at the printed precision are all bold. Four-GNN averages and SDs are in Table 5.

Configuration	LI-Small		HI-Small		LI-Medium	
	EdgeSAGE	EdgeGIN	EdgeSAGE	EdgeGIN	EdgeSAGE	EdgeGIN
Real (TRTR)	0.333	0.241	0.502	0.484	0.110	0.084
Learned DBA-KL (terminal)	0.329	0.201	0.538	**0.488**	**0.090**	**0.067**
Calibrated DBA-KL	0.329	0.201	0.538	**0.488**	**0.090**	**0.067**
Fixed-topology DBA-KL	0.330	0.197	0.523	0.489	0.093	0.070
DBA-KL/projected EDM (3 rounds)	**0.334**	**0.202**	**0.545**	0.484	**0.090**	**0.067**
ECI-DBA (1 round)	0.316	0.194	0.526	0.473	0.080	0.059
ECI-DBA (3 rounds)	0.309	0.194	0.518	0.468	0.073	0.053
PCFM-log	0.317	0.203	0.525	0.485	0.079	0.060
ECI-naive (equality only)	0.304	0.194	0.517	0.446	0.060	0.047
PCFM-naive (equality only)	0.304	0.194	0.517	0.446	0.060	0.047
HARPOON (raw output)	0.213	0.104	0.365	0.236	0.069	0.048
GDSS (raw output)	0.152	0.064	0.379	0.217	0.058	0.043

Table 2: AP of *Structure + diffusion, without flow completion* (`rq2_no_flow`); mean ± reported SD. The final row retains the supplied SD of seedwise four-GNN averages, not an average of model SDs. Relay-invalidity measurements are in Appendix D.5.

Evaluator	LI-Small	HI-Small	LI-Medium
EdgeSAGE	0.344723 ± 0.065006	0.554808 ± 0.026958	0.111609 ± 0.088776
EdgeGCN	0.201739 ± 0.027527	0.396518 ± 0.018536	0.017859 ± 0.003444
EdgeGAT	0.071676 ± 0.013579	0.234388 ± 0.031184	0.036052 ± 0.015281
EdgeGIN	0.222186 ± 0.024598	0.485055 ± 0.036248	0.082627 ± 0.057412
Four-GNN average	0.210081 ± 0.017115	0.417692 ± 0.011385	0.062037 ± 0.032998

Results. Every DBA-KL configuration exceeds every ECI-DBA and PCFM-log configuration in four-GNN average AP on every dataset (Table 5), and terminal Learned does so on every detector seed. The topology-changing variants improve on ECI-DBA by 7.9%, 5.7%, and 23.5%, and on PCFM-log by 3.0%, 2.6%, and 12.7%. The margin over the best adaptation is widest on LI-Medium, at 17.3% against 3.0% and 2.6%. The single exception in Table 1 is LI-Small EdgeGIN, where PCFM-log leads by 0.2%. Terminal Learned is 2.3% above real-data training on HI-Small, and 6.9% and 17.5% below it on LI-Small and LI-Medium. Both external generators trail every DBA-KL configuration and every adaptation on every evaluator of every dataset (Tables 6–8); in four-GNN average AP, HARPOON is 44.1%, 38.7%, and 25.6% below terminal Learned and GDSS 62.5%, 39.7%, and 36.5%. Absolute AP is not comparable across datasets (Appendix D.3).

4.3 RQ2: Does predictive utility separate valid from invalid generation?

Setting. We add *Structure + diffusion, without flow completion* (`rq2_no_flow`) to the native component ablations. It keeps the topology menu, learned structural scores, diffusion-generated attributes, pinned observations, and temporal completion. It removes quantity correction, flow-feasibility filtering, and quantity-distortion weighting. The replacement task and detector settings are unchanged.

Validity and accuracy diverge. This ablation attains higher four-GNN mean AP than terminal Learned DBA-KL on every dataset: 8.73%, 2.87%, and 23.11% relative to terminal Learned and 1.26%, 5.24%, and 1.54% relative to Real/TRTR. Yet all 610/610 audited relay-containing outputs violate at least one relay equation (Appendix D.5). The 485 relay-free outputs are N/A, not conservation successes. This experiment therefore evaluates validity, not predictive gains from com-

Table 3: Saved-skeleton audit, one stored draw per offered source. Valid: all seven predicates pass (share of offered sources); edited: fixed-account typed incidence differs from the source (share of emitted); valid new: valid and non-isomorphic to the source (share of offered); relay-free: share of emitted. Within these completion arms, invalid graphs fail only the closed account-vocabulary predicate. Per-configuration counts for all ten flow-corrected arms are in the supplied `results/rq3/audit/` exports.

Dataset	Arm	Emitted	Valid	Edited	Valid new	Relay-free
LI-Small	Fixed-topology	146/159	146 (91.8%)	0 (0.0%)	0 (0.0%)	37 (25.3%)
	Learned	148/159	119 (74.8%)	89 (60.1%)	55 (34.6%)	71 (48.0%)
HI-Small	Fixed-topology	403/459	403 (87.8%)	0 (0.0%)	0 (0.0%)	120 (29.8%)
	Learned	411/459	321 (69.9%)	293 (71.3%)	182 (39.7%)	196 (47.7%)
LI-Medium	Fixed-topology	481/544	481 (88.4%)	0 (0.0%)	0 (0.0%)	160 (33.3%)
	Learned	485/544	382 (70.2%)	353 (72.8%)	224 (41.2%)	260 (53.6%)

pletion. A generator can score better downstream while producing quantitatively invalid output on essentially every constrained example. Downstream utility on its own is a poor guide to constrained graph generation, and validity has to be measured separately.

Topology-enabled generation retains predictive utility. Permitting alternative connectivity is nonetheless compatible with retaining predictive utility here, and it need not incur an aggregate penalty. Alternative structures can exercise relationships that fixed-topology attribute generation cannot reach. The result concerns the pipeline as a whole; it is not a topology-only causal effect or a claim of statistical non-inferiority.

Existing ablations. With DBA-KL retained, editing and selection improve the Small-dataset averages but not LI-Medium, where fixed topology remains the best flow-corrected native arm. Topology editing thus trades some predictive similarity for additional valid structural coverage. Extra rounds and calibration do not consistently improve AP, and ECI-DBA uses DBA-KL internally, so that comparison concerns the schedule. We ran no cost-weighting-only control, no uniform-policy arm, and no common-source/count control. The ablation above changes correction, feasibility filtering, and weighting jointly, so it cannot isolate the effect of the cost term (Section 4.4).

4.4 RQ3: Are generated scenarios valid and structurally diverse?

Setting. We audit the ten saved flow-corrected configurations, using one stored draw per offered source (159, 459, and 544). Each output is checked on the modeled event set against the seven validity predicates of Appendix D.6. Relay-free graphs are not applicable to conservation.

Validity. Relay balance holds within tolerance for every emitted graph of every completion arm, whereas 88–97% of the original records violate the zero-relay scenario model, a construction rule we impose on generated outputs that original case records need not satisfy. Positivity, pins, time order, and source family also pass. The closed observed-account vocabulary is the only predicate that fails, for 29, 90, and 103 edited graphs of terminal Learned. The two external generators, which trail every DBA-KL configuration and every adaptation in Table 1, also violate the quantitative properties outright, with relay-balance errors of 4.2×10^7 to 8.0×10^9 common units and 0.35–2.32% negative quantities (Appendix D.6).

Structural variation. Terminal Learned edits the topology of 60%, 71%, and 73% of its emitted graphs, and 35%, 40%, and 41% of the sources receive a valid topology with no identity-preserving isomorphism to the source. Across the ten generated configurations, 45–56% of the sources receive at least two distinct valid structures. Coverage is incomplete: the grammar cannot represent the source skeleton for 5, 26, and 33 offered sources, and further omissions are boundary infeasibility or non-convergence. Among candidates with complete diagnostics, 88.9–93.8% are accepted under positivity-aware completion, against 82.5–88.4% for the equality-only controls. This coverage comes at the cost of predictive similarity (Section 4.3).

Relay-free outputs. A relay-free candidate has no conservation equation and therefore zero completion cost, so cost-weighted selection can favour structures for which there is nothing to conserve. Terminal Learned is relay-free for 48.0%, 47.7%, and 53.6% of its emitted graphs, against 25.3%, 29.8%, and 33.3% for fixed topology (Table 3). The unweighted ablation of Section 4.3 is *less* relay-free (42.5%, 42.0%, 46.8%) and emits more, which is consistent with the cost term encouraging relay-free candidates.

4.5 RQ4: What are the computational costs and scaling limits?

Setting. We time the completion step alone. The same frozen (A, b, q) is solved by native DBA-KL at production tolerances, by the same-KL Newton solve of Appendix E, by the fixed-reference PCFM-log projection, and by the positive QP. The problems are constructed with $r \in \{1, 2, 4, 8\}$ relays in five structures, 25% pinned events, and log-normal proposal noise of SD 0.5 or 1.5, with 50 instances per cell (Appendix D.7). All times come from single-threaded Python reference implementations and exclude the sampler and denoiser; Table 10 in Appendix D.7 gives the per-cell values.

Runtime. DBA-KL is not uniformly faster than the alternatives. Its main advantage is that each relay update preserves positivity in closed form and optimizes a fixed-reference KL objective, whose required/excess split puts every completion on a common scale. The measured speed depends on structure and says nothing about whole-generation runtime. Every DBA-KL call converged to the Newton optimum, with excess KL below 5×10^{-16} per unit of source scale. PCFM-log failed 1 and 2 of 50 calls on layered four- and eight-relay stacks at SD 1.5. On decoupled structures DBA-KL needs one sweep and 0.41–0.65 ms per call, 3–5 times less than Newton and 17–78 times less than PCFM-log (Table 10). The terminal schedule also solves once per candidate.

Scaling limit. Relay coupling is the limiting factor. Chains of 2, 4, and 8 relays need 14, 60, and 106 median sweeps (maximum 330), so a chain-8 call costs 24.3 ms (P95 57–66 ms) against 1.3–2.9 ms for Newton. Newton is therefore 3–11 times cheaper on chains of four or more relays, and DBA-KL is the slower solver in 10 of the 38 cells, all with at least 14 sweeps. Both reach the same optimum, so a sweep budget with a Newton fallback would bound the cost without changing the accounting.

5 Limitations

All evidence comes from one simulator family (three IBM-AML subsets), single fitted generators, and stored draws; detector-seed SDs do not measure generation uncertainty, and no second domain was evaluated. The no-completion ablation does not isolate topology's effect on AP, and the relay-free preference is reported without identifying the effect of the cost term. Component-level evaluation of the structural policy, proposal prior, and menu-size sensitivity remains incomplete. Relay validity is model-specific, not task-label validity; timings exclude full generation (Appendix H).

6 Conclusion

We presented topology-conditioned feasible-set generation for continuous-time multigraphs: each structural proposal induces its own constraint system, and DBA-KL supplies an exact positive completion whose distortion is fed back into selection, with positivity-preserving updates and a required/excess split. On three AML benchmarks from one simulator family, every DBA-KL configuration exceeds the ECI-DBA and PCFM-log adaptations in four-GNN average AP, and the terminal variant exceeds real-data training on HI-Small. Most importantly, removing flow completion raises AP on every dataset while making every audited relay-containing output invalid; predictive utility is a poor guide to validity.

AI DISCLOSURE

In this work, we used large language models (LLMs) to assist with drafting sections of the manuscript and polishing the writing for clarity, readability, and organization. We also used LLMs for information retrieval and literature discovery, including identifying potentially relevant related work. Beyond writing and literature support, LLMs assisted with developing and writing proofs of mathematical claims and with generating the synthetic datasets used in RQ4, the controlled completion-cost and scalability study (Appendix D.7).

The authors take full responsibility for the final manuscript and all associated research artifacts, including AI-assisted text, mathematical arguments, synthetic data, and the conclusions drawn from the experiments. This responsibility includes checking factual accuracy and citations, assessing the correctness of mathematical claims and proofs, and validating the synthetic data and experimental procedures.

REFERENCES

Erik Altman, Jovan Blanuša, Luc von Niederhäusern, Béni Egressy, Andreea Anghel, and Kubilay Atasu. Realistic synthetic financial transactions for anti-money laundering models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29851–29874, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5f38404edff6f3f642d6fa5892479c42-Abstract-Datasets_and_Benchmarks.html.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. doi: 10.1016/0041-5553(67)90040-7. URL https://www.sciencedirect.com/science/article/pii/0041555367900407.

Chaoran Cheng, Boran Han, Danielle Maddix, Abdul Fatir Ansari, Andrew Stuart, Michael W. Mahoney, and Bernie Wang. Gradient-free generation for hard-constrained systems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f9d7d6c695bc983fcfb5b70a5fbdfd2f-Abstract-Conference.html.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a28af221f2f70be183afc16797a56b91-Abstract-Conference.html.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IuXR1CCrSi.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0dc91de822b71c66a7f54fa121d8cbb9-Abstract-Datasets_and_Benchmarks.html.

Derek Hansen, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, and Michael W. Mahoney. Learning physical models that can respect conservation laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12469–12510. PMLR, 2023. URL https://proceedings.mlr.press/v202/hansen23b.html.

Ryien Hosseini, Filippo Simini, Venkatram Vishwanath, and Henry Hoffmann. A deep probabilistic framework for continuous time dynamic graph generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17249–17257, 2025. doi: 10.1609/aaai.v39i16.33896. URL https://ojs.aaai.org/index.php/AAAI/article/view/33896.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10362–10383, 2022. URL https://proceedings.mlr.press/v162/jo22a.html.

Jaehyeong Jo, Dongki Kim, and Sung Ju Hwang. Graph generation with diffusion mixture. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22371–22405, 2024. URL https://proceedings.mlr.press/v235/jo24b.html.

Nuowen Kan, Sa Yan, Junni Zou, Wenrui Dai, Xing Gao, Chenglin Li, and Hongkai Xiong. GDPlan: Generative network planning via graph diffusion model. *IEEE Transactions on Networking*, 2025. URL https://ieeexplore.ieee.org/document/10858184/.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022. URL https://arxiv.org/abs/2206.00364.

Ana Lucic, Maartje A. ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4499–4511, 2022. URL https://proceedings.mlr.press/v151/lucic22a.html.

Manuel Madeira, Clément Vignac, Dorina Thanou, and Pascal Frossard. Generative modelling of structurally constrained graphs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f82385b8804009f9a81e1a30f1ff14e3-Abstract-Conference.html.

Antoine Marot, Benjamin Donnot, Gabriel Dulac-Arnold, Adrian Kelly, Aidan O'Sullivan, Jan Viebahn, Mariette Awad, Isabelle Guyon, Patrick Panciatici, and Camilo Romero. Learning to run a power network challenge: a retrospective analysis. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 112–132, 2021. URL https://proceedings.mlr.press/v133/marot21a.html.

Yiming Qin, Manuel Madeira, Dorina Thanou, and Pascal Frossard. DeFoG: Discrete flow matching for graph generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50269–50326. PMLR, 2025. URL https://proceedings.mlr.press/v267/qin25d.html.

Aditya Shankar, Yuandou Wang, Rihan Hai, and Lydia Y. Chen. Harpoon: Generalised manifold guidance for conditional tabular diffusion. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=G5g6tDg1ZE.

Yuxing Tian, Aiwen Jiang, Qi Huang, Jian Guo, and Yiyan Qi. Latent diffusion-based data augmentation for continuous-time dynamic graph model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2900–2911, 2024. doi: 10.1145/3637528.3671863. URL https://dl.acm.org/doi/10.1145/3637528.3671863. The arXiv version is titled "Latent Conditional Diffusion-based Data Augmentation for Continuous-Time Dynamic Graph Model".

Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gómez-Bombarelli, and Christopher Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints, 2025. URL https://arxiv.org/abs/2506.04171.

Fedor Velikonivtsev, Mikhail Mironov, and Liudmila Prokhorenkova. Challenges of generating structurally diverse graphs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ac807c9b296964409b277369e55621a-Abstract-Conference.html.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14734.

Yuwei Xu and Chenhao Ma. MoDiff: Graph generation with motif-aware diffusion model. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3711896.3737053. URL https://dl.acm.org/doi/10.1145/3711896.3737053.

Xuesong Zhang, Haifeng Sun, Qi Qi, Shengkuan Li, Yuhao Li, Tianyi Kou, Zirui Zhuang, Bo He, Jianxin Liao, and Jingyu Wang. Hard-constrained graph generation with discrete-projection diffusion. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cbtykHVWX9.

Cai Zhou, Xiyuan Wang, and Muhan Zhang. Unifying generation and prediction on graphs with latent graph diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 61963–61999, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/718d02a76d69686a36eccc8cde3e6a41-Abstract-Conference.html.

Dawei Zhou, Lecheng Zheng, Jiawei Han, and Jingrui He. A data-driven graph generative model for temporal interaction networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 401–411. Association for Computing Machinery, 2020. doi: 10.1145/3394486.3403082. URL https://www.kdd.org/kdd2020/accepted-papers/view/a-data-driven-graph-generative-model-for-temporal-interaction-networks.html.

A TOPOLOGY-DEPENDENT FEASIBLE SPACES AND STRUCTURAL COVERAGE

This appendix formalizes the topology-dependent generative domain referred to in Sections 2.2 and 3.

Let $\mathcal{S}_{\mathcal{B}}$ denote the candidate structures satisfying the structural predicates and observation boundary. Each candidate induces a continuous feasible set,

$$\mathcal{F}_{\mathcal{B}}(S) = \{(v,\tau) : v > 0,\ A_R(S)v = 0,\ (v,\tau)_{P_c} = (\bar{v},\bar{\tau})_{P_c},\ \tau \in \mathcal{T}_{\mathcal{B}}(S)\}\,. \tag{6}$$

Here $\mathcal{T}_{\mathcal{B}}(S)$ contains the temporal constraints defined in Section 3.4. The conditional generative domain is the disjoint union

$$\Omega_{\mathcal{B}} = \bigsqcup_{S \in \mathcal{S}_{\mathcal{B}}} \{S\} \times \mathcal{F}_{\mathcal{B}}(S). \tag{7}$$

Structural edits can change the number of free variables, the relay set, and the feasible region. Structural admissibility and continuous feasibility must therefore be treated separately. The implementation explores a bounded one-edit neighborhood within this domain.

Conventions of the balance model. The relay set $V_R(S)$ of Section 2.2 is recomputed for each candidate from its modeled events; external context incidence does not create relays, and pure sources and sinks have no relay-balance equations. A valid skeleton with $V_R(S) = \varnothing$ has no relay-balance obligation and is reported as not applicable rather than counted as a nontrivial conservation success. In the signed incidence $A_R(S)$, parallel events remain separate columns, and a self-interaction, if represented among fixed events, cancels in the signed equation. The right-hand side $b_S = -A_{R,P_c}(S)\bar{v}_{P_c}$ is determined solely by the fixed modeled quantities; context-only flows do not enter the system.

Structural coverage and task semantics. The topology projection of $\Omega_{\mathcal{B}}$ consists exactly of those $S \in \mathcal{S}_{\mathcal{B}}$ for which $\mathcal{F}_{\mathcal{B}}(S) \neq \varnothing$. Fixed-topology generation explores at most one such topology per source, regardless of how many continuous variants it draws. Structural novelty must be distinguished from consistent identifier renaming and from changes in how a skeleton attaches to its observed context. A coverage comparison therefore specifies which identities, boundary relations, edge directions, and multiplicities are preserved by its equivalence relation. Membership in $\mathcal{F}_{\mathcal{B}}(S)$ establishes the stated quantitative and temporal conditions, not a downstream task label. Task labels and any additional application-specific validity requirements are established separately.

B DERIVATIONS FOR THE COMPLETION AND CALIBRATION OBJECTIVES

The following statements are consequences of the finite-dimensional objectives used in Section 3.3 and Appendix C.7. They do not assert that every structural candidate is feasible or provide a global iteration bound for the finite numerical implementation.

Proposition 1 (Existence, positivity, and uniqueness of quantity completion)**.** *Let $\widehat{v} \in \mathbb{R}^m_{>0}$ and suppose that there exists $z^+ \in \mathbb{R}^m_{>0}$ with $Az^+ = b$. Then minimizing $D^+_{\mathrm{KL}}(z\|\widehat{v})$ subject to $z > 0$ and $Az = b$ has a unique minimizer. It satisfies $\log(z^\star/\widehat{v}) + A^\top\lambda^\star = 0$ for some multiplier. No full-row-rank assumption on A is required for uniqueness of $z^\star$.*

Proof. Extend $z_i \log(z_i/\widehat{v}_i)$ continuously to zero with value zero. The resulting divergence is continuous and coercive on the nonnegative orthant: it tends to infinity as any coordinate tends to infinity. The nonnegative feasible set is closed and nonempty, so a minimizer exists. Suppose a minimizer z had a zero coordinate. For $z(t) = (1-t)z + tz^+$, the change contributed by the zero coordinates is $t \log t \sum_{i:z_i=0} z_i^+ + O(t)$, whereas the other coordinates contribute $O(t)$. The total is negative for sufficiently small positive t, a contradiction. The minimizer is therefore strictly positive. The Hessian of the divergence there is $\mathrm{diag}(1/z_i) \gg 0$, so strict convexity gives uniqueness. At an interior affine minimizer, the gradient is orthogonal to $\ker A$ and hence belongs to $\mathrm{range}\, A^\top$, giving the displayed stationarity equation. Redundant rows may make multipliers nonunique without changing the primal minimizer. □

Proposition 2 (Exact one-row dual update)**.** *If row a has current positive free totals $I, O > 0$, the update in Eq.* (15) *uniquely maximizes the dual along coordinate a and makes that row satisfy $Az = b$ immediately after the update in exact arithmetic.*

Proof. Up to a constant, the one-dimensional dual is $-Ie^{-\delta} - Oe^{\delta} - b_a\delta$. Its derivative is $Ie^{-\delta} - Oe^{\delta} - b_a$, and its second derivative is strictly negative. The derivative tends to $+\infty$ and $-\infty$ at the two ends of the real line, respectively. Substituting $s = e^{\delta} > 0$ gives $Os^2 + b_a s - I = 0$. Its positive root yields Eq. (15); the two branches use algebraically equivalent forms to avoid cancellation. The derivative equation is precisely the row's balance equation. Other rows can change because they share event quantities. □

Proposition 3 (Row-wise Pythagorean identity of an exact relay update)**.** *Let $z > 0$ and let $z'_i = z_i \exp(-\delta A_{a,i})$ be an exact update satisfying $A_{a,:}z' = b_a$. For every $u > 0$ with $A_{a,:}u = b_a$,*

$$D^+_{\mathrm{KL}}(u\|z) = D^+_{\mathrm{KL}}(u\|z') + D^+_{\mathrm{KL}}(z'\|z).$$

Proof. Expanding the three divergences leaves the cross term

$$(u - z')^\top \log(z'/z) = -\delta A_{a,:}(u - z') = 0.$$

Every globally feasible u satisfies this row equation, which gives the identity along the whole cyclic sequence. Only the updated row is balanced immediately; the remaining rows follow later in the sweep. □

Proposition 4 (Convergence of exact cyclic relay updates)**.** *Suppose $q > 0$ and $Az^+ = b$ for some $z^+ > 0$. Starting from $z^0 = q$ and $\lambda^0 = 0$, visit each nonzero relay row in a fixed finite cyclic order and perform the exact update in Eq.* (4)*. Consistent zero rows are skipped. Then the primal iterates remain positive and converge to the unique minimizer $z^\star$ of Eq.* (2)*. This statement does not supply a uniform finite sweep bound.*

Proof. Strict positive feasibility ensures that each nonzero row has a positive feasible assignment. Thus its update exists: two-sided rows use the positive quadratic root, and one-sided rows have the required sign of b_a and a finite logarithmic update. Positive finite coordinates remain positive. By Proposition 3,

$$D^+_{\mathrm{KL}}(z^+\|z^{k+1}) + D^+_{\mathrm{KL}}(z^{k+1}\|z^k) = D^+_{\mathrm{KL}}(z^+\|z^k).$$

For fixed $z^+ > 0$, each coordinate contribution to $D^+_{\mathrm{KL}}(z^+\|z)$ diverges when that coordinate of z tends to zero or infinity. The iterates therefore lie in a compact subset of the positive orthant. Telescoping the identity shows that $\sum_k D^+_{\mathrm{KL}}(z^{k+1}\|z^k) < \infty$. On this compact set the entropy Hessian is uniformly positive definite, so $\|z^{k+1} - z^k\| \to 0$.

Each row is satisfied immediately after its visit. Since a sweep contains finitely many updates and successive differences vanish, every row residual converges to zero, including at sweep boundaries. Every cluster point $\bar{z}$ is consequently positive and satisfies $A\bar{z} = b$. The maintained parameterization gives $\log(z^k/q) \in \mathrm{range}(A^\top)$ for all k. This is a closed linear subspace, and the logarithm is continuous on the compact set, so $\log(\bar{z}/q) \in \mathrm{range}(A^\top)$. Hence $\bar{z}$ satisfies the affine first-order optimality condition. Strict convexity identifies it with $z^\star$. All cluster points coincide, which proves convergence of the whole primal sequence. No row-rank assumption or convergence of a particular multiplier representative is needed. □

Acceptance versus convergence. The actual routine maintains the exponential dual parameterization, checks stationarity, and accepts only a numerically feasible positive iterate. A successful final graph audit certifies the reported relay tolerances, not a universal exact-arithmetic guarantee. A finite sweep limit or dynamic-range failure is not by itself a proof that no strictly positive feasible point exists. The stationarity residual is checked in log coordinates at 10^{-7}; the reported dual-gap diagnostic is not used as a universal optimality stopping threshold.

Proposition 5 (Class-mass preservation and total KL budget of the calibration correction)**.** *Fix a completed candidate table with base probability p and partition c. The correction in Eq.* (17) *preserves $p(c)$ for every class. If every conditional divergence is at most δ_{KL}, then*

$$D_{\mathrm{KL}}(p_\beta\|p) = \sum_c p(c) D_{\mathrm{KL}}(p_\beta(\cdot \mid c)\|p(\cdot \mid c)) \le \delta_{\mathrm{KL}}.$$

This statement is conditional on the fixed candidate table and does not bound changes in downstream classification performance.

Proof. Within-class normalization sums to one, so summing $p_\beta(e)$ over $e \in c$ yields $p(c)$. In the log probability ratio, the identical class masses cancel. Summing the remaining conditional log ratios gives the displayed identity and bound. In finite arithmetic, the implementation reports both the class-mass error and maximum conditional KL. □

Why bisection is applicable. For one class with positive base probabilities and score s_e, let $Z(t) = \sum_e p_e e^{ts_e}$ and $q_t(e) = p_e e^{ts_e}/Z(t)$. Then $K(t) = D_{\mathrm{KL}}(q_t \| p) = t\mathbb{E}_{q_t} s - \log Z(t)$ and $K'(t) = t \operatorname{Var}_{q_t}(s) \geq 0$ for $t \geq 0$. This justifies reducing t by bisection to meet the conditional budget; it does not justify forcing a correction on when the held-out gate declines it.

C Component algorithms and implementation details

This appendix records the component algorithms, training rules, and numerical conventions underlying Section 3. It separates these details from the fixed-proposal quantity objective and its exact distortion accounting.

C.1 Fitting partitions, units, and generation pins

The modules are fitted separately on parent-grouped outer-training data. Approximately 60% of parents fit the policy, prior, and their preprocessing statistics; 20% fit optional selection calibration; and 20% independently gate its use. Sibling components stay together. The prior and policy are not refitted on calibration or gate parents, and outer-test cases fit none of these components. Generation uses the outer-training sources with all fitted components frozen. Quantity conversion uses a positive table estimated on the fitting split and then frozen; conservation and distortion use this same common-unit representation. In the financial implementation, a native quantity is divided by its recorded conversion rate. This convention defines the modeled quantity; it is not an independently measured physical exchange law. Algorithm 1 summarizes inference.

Algorithm 1 Main terminal generator for one source

1. Freeze $\mathcal{B}$, observations on P_c, and fitted components; construct $\mathcal{M}$ and compute policy scores.
2. Generate and freeze one proposal per candidate; apply DBA-KL and temporal completion with fixed observations unchanged.
3. Validate completed graphs and retain $\mathcal{M}_{\mathrm{ok}}$ with numerical costs; record exclusions and omit sources with an empty menu.
4. Sample by Eq. (5), using the optional calibration only when its held-out gate passes.
5. Emit the completed graph and verify its stored boundary and invariants.

We use the event partition in Section 2.2: pinned pattern events are P, all fixed modeled events are $P_c = P \cup E^{\mathrm{att}}$, and candidate S_e has free events $F_e = E^{\mathrm{pat}}_{S_e} \setminus P$. Generation pins combine explicit pins with a reproducible subset of $\lfloor \rho |E^{\mathrm{pat}}_0| \rfloor$ pattern events, with default $\rho = 0.25$; the union can be larger and small patterns can have no sampled pin. All observations on P_c remain fixed. Context-only events condition generation but neither create relays nor enter the balance equations. In the graph representation, modeled events are core transactions; core events marked `tier=embed` are the fixed attachments, and the other core events are pattern events. Non-core context/halo events are outside the balance model. Legacy role or gauge annotations do not remove an incidence-defined relay.

C.2 Structural menu and policy fitting

For each source, a typed grammar constructs a finite menu $\mathcal{M}(S_0, \mathcal{B})$ containing the no-edit state and family-compatible insertions, deletions, rewiring, subdivision, contraction, and intermediary replacement. Edits act on unpinned pattern events, preserve all events in P_c, and use entity identities observed in the source or its context. Deterministic round-robin enumeration and endpoint-multiset deduplication retain at most K candidates. Categorical metadata follows the dataset convention,

without a separate diffusion process. The grammar covers fans, bipartite patterns, acyclic stacks, directed cycles, gather–scatter, scatter–gather, and non-self-loop general patterns. Its bounded vocabulary and menu restrict the explored support; they do not assert that every feasible topology is represented. The default alternative-identity vocabulary has at most 24 entries and the menu has $K = 64$ states. Both identity and descriptor ordering are reproducible; the vocabulary cap does not restrict entities already present in the source. For the financial adapter, retained events keep their currency labels and added events receive the source pattern's modal currency. Source-median quantities and times initialize new-event placeholders, not the final free attributes.

A standardized feature vector $\phi(S_0, e)$ describes edit type, event counts, width changes, degrees, repeated pairs, connectivity, endpoint-group compatibility, and family interactions. Amounts and timestamps are not policy features. The log-linear structural policy is given by Eq. (1). Training reconstructs original topologies from synthetically corrupted skeletons. When nontrivial corruptions are available, we use the unchanged skeleton with probability 0.2 and otherwise sample an edit type and an instance uniformly. A reconstruction target must occur in the corrupted skeleton's admissible menu; absent targets are recorded as coverage failures. For parents $\mathcal{H}_+$ with nonempty reconstructible task sets $\mathcal{D}_h$, we minimize the parent-balanced objective

$$\mathcal{L}_{\text{struct}} = -\frac{1}{|\mathcal{H}_+|} \sum_{h \in \mathcal{H}_+} \frac{1}{|\mathcal{D}_h|} \sum_{j \in \mathcal{D}_h} \log q_\psi(e_j^\star \mid \widetilde{S}_j, \mathcal{B}_j) + \frac{\lambda_\psi}{2} \|\psi\|_2^2. \tag{8}$$

Each parent therefore has equal total loss weight. This learns conditional structural reconstruction, not an observed temporal mutation law. Feature standardization uses fitting tasks only. Targets absent from the admissible menu are recorded as coverage failures. A separate menu-recall or reconstruction-accuracy statistic and an ablation of the menu-size budget K remain outside this study.

C.3 Boundary-conditioned EDM: training and sampling

Each modeled event has two continuous coordinates: standardized log quantity and a scaled logit of normalized time. A shared two-hidden-layer SiLU MLP uses EDM preconditioning (Karras et al., 2022). Its conditioning $\mathbf{c}_i$ contains family, categorical metadata, endpoint roles, degrees, graph counts, and context summaries. A boundary block B_i adds pin masks, prescribed clean coordinates, pinned-flow aggregates, and context quantities. The backbone is row-wise; shared summaries and constrained completion couple the events. Conditioning on the observed boundary supplies information during denoising, whereas the later completion operator enforces the hard equations. These are complementary roles, not interchangeable mechanisms. Event time τ is distinct from the artificial noise coordinate σ of the diffusion; connectivity stays fixed during denoising even though event times are continuous.

Encoding and decoding. For the graph S supplied to the encoder, let $t^{\text{core}}_{\min}(S)$ and $t^{\text{core}}_{\max}(S)$ be the extrema over its modeled core events, including fixed core attachments. Define

$$\begin{aligned} f_i(S) &= \operatorname{clip}\left(\frac{\tau_i - t^{\text{core}}_{\min}(S)}{\max\{t^{\text{core}}_{\max}(S) - t^{\text{core}}_{\min}(S), 1\}}, 10^{-3}, 1 - 10^{-3} \right), \\ x_{0,i} &= \left[\frac{\log \max(v_i, 10^{-9}) - \mu}{s}, \ \frac{1}{4} \log \frac{f_i(S)}{1 - f_i(S)} \right]. \end{aligned} \tag{9}$$

The log-quantity statistics μ and s are estimated from the fitting subset. At inference, $x_{0,i}$ supplies observed coordinates for fixed events; free coordinates are initialized from noise rather than their placeholder values.

The free-time decoder uses the original source pattern window, not the encoder's current core-event window. With $[\ell_0, u_0]$ denoting that source window, the decoded proposals are

$$\begin{aligned} \widehat{v}_i &= \exp\big(\operatorname{clip}(s x_{i,1} + \mu, -20, 35)\big), \\ \widehat{\tau}_i &= \ell_0 + (u_0 - \ell_0) \operatorname{sigmoid}(4 x_{i,2}). \end{aligned} \tag{10}$$

Clipping counts are recorded, and fixed physical quantities and timestamps are restored explicitly after decoding. These encodings need not be an inverse pair under one normalization.

Training. Training perturbs observed coordinates as $x_{\sigma,i} = x_{0,i} + \sigma_i \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, I)$ and $\log \sigma_i \sim \mathcal{N}(-1.2, 1.2^2)$. For $\sigma_{\text{data}} = 1$, EDM preconditioning gives

$$D_\theta(x, \sigma, \mathbf{c}, B) = \frac{x}{1+\sigma^2} + \frac{\sigma}{\sqrt{1+\sigma^2}} F_\theta\left(\frac{x}{\sqrt{1+\sigma^2}}, \frac{\log \sigma}{4}, \mathbf{c}, B\right). \tag{11}$$

For M event rows and $\widehat{x}_{0,i} = D_\theta(x_{\sigma,i}, \sigma_i, \mathbf{c}_i, B_i)$, the loss is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\frac{1}{2M} \sum_{i=1}^{M} \frac{1+\sigma_i^2}{\sigma_i^2} \sum_{j=1}^{2} \omega_{i,j} (\widehat{x}_{0,i,j} - x_{0,i,j})^2\right]. \tag{12}$$

Family-dependent masks expose observed coordinates; pinned coordinates are downweighted to 0.25 by default. Conditioning dropout, applied with probability 0.1 per sampled motif, hides the boundary block and context-quantity summary and uses an all-free mask with unit loss weights. Targets remain original observations. Motifs are sampled and their event rows concatenated; this event-averaged objective is distinct from the parent-balanced policy loss.

For each candidate, conditioning is recomputed and Gaussian noise is keyed by source, event identity, and generation seed. The generation-time fixed mask covers P_c, not only P. Fixed coordinates start at their observed encoding plus scaled noise. With $\widetilde{D}_\theta$ overwriting fixed clean predictions on P_c, generation uses Euler sampling,

$$x_{k+1} = x_k + (\sigma_{k+1} - \sigma_k) \frac{x_k - \widetilde{D}_\theta(x_k, \sigma_k, \mathbf{c}, B)}{\sigma_k}, \tag{13}$$

along a decreasing schedule ending at zero. The default uses 40 positive noise levels, exponent 7, and endpoints 80 and 0.002. At the final zero-noise transition the state equals the clean prediction. Free amounts are decoded by exponentiation and times into the source pattern window; fixed physical observations are then restored explicitly. The resulting proposal $(\widehat{v}_e, \widehat{\tau}_e)$ is frozen throughout completion and selection. The implementation uses a core-window time encoder and a source-window time decoder; these need not form an inverse pair under one normalization.

C.4 Stable relay updates and numerical acceptance

DBA-KL maintains the exponential dual parameterization of Eq. (2) and cyclically maximizes its dual,

$$z(\lambda) = q \odot \exp(-A^\top \lambda), \qquad g(\lambda) = \sum_i (q_i - z_i(\lambda)) - b^\top \lambda, \tag{14}$$

starting at $z^0 = q$, $\lambda^0 = 0$; the row update of Eq. (4) is the exact maximization of g along coordinate a (Proposition 2). For Eq. (4), when $I, O > 0$, the scalar e^δ is the positive root of $Ot^2 + b_a t - I = 0$. The implementation uses

$$\delta = \begin{cases} \log(2I) - \log\left(b_a + \sqrt{b_a^2 + 4IO}\right), & b_a \geq 0, \\ \log\left(-b_a + \sqrt{b_a^2 + 4IO}\right) - \log(2O), & b_a < 0. \end{cases} \tag{15}$$

When $b_a = 0$, $\delta = \frac{1}{2}\log(I/O)$ and both updated totals equal $\sqrt{IO}$. A row with only incoming free quantities requires $b_a > 0$ and uses $\delta = \log(I/b_a)$; one with only outgoing free quantities requires $b_a < 0$ and uses $\delta = \log(-b_a/O)$. Rows with no free coefficients are checked directly. Shared events couple relays, so all equations are rechecked after each sweep.

Acceptance. The numerical solver has a bounded sweep budget and returns a positive finite $\widetilde{v}_{F_e}$ only when its stationarity check and every relay tolerance pass: $|I_a^{S_e} - O_a^{S_e}| \leq \epsilon_{\text{abs}} + \epsilon_{\text{rel}} \max(I_a^{S_e}, O_a^{S_e})$. The defaults are 10^{-10} absolute, 10^{-8} relative, and 10^{-7} for stationarity in log coordinates. The relay totals include fixed modeled quantities. An independent audit checks the actual stored graph. A small balance residual on its own is weaker than an exact KL optimum: finite-precision accounting must distinguish numerical deviations from the ideal identity of Eq. (3) (their measured size at native and tight tolerances is in Appendix E). Numerical failures are recorded, not treated as proofs of infeasibility. Neither ideal convergence nor analytic local updates guarantee completion within a fixed budget or universal runtime superiority.

C.5 Accounting derivation and allocation invariants

Propositions 1–4 in Appendix B establish the reference solution, exact row updates, and ideal cyclic convergence under strict positive feasibility. The following expansion explains the distortion accounting used in Section 3.3.

Optimality gives $\log(z^\star/q) = -A^\top \lambda^\star$. For any positive u with $Au = b$,

$$\begin{aligned} &D^+_{\mathrm{KL}}(u\|q) - D^+_{\mathrm{KL}}(z^\star\|q) - D^+_{\mathrm{KL}}(u\|z^\star) \\ &\quad = (u - z^\star)^\top \log(z^\star/q) = -\lambda^{\star\top} A(u - z^\star) = 0. \end{aligned}$$

This proves Eq. (3) for the same fixed (A, b, q). It applies to any positive feasible output, not only DBA-KL's output; an accurate solver of the same KL objective shares the reference. The equality does not assert a causal decomposition of individual transaction changes or extend to the inequality-constrained temporal projection.

The maintained form $z^k = q \odot \exp(-A^\top \lambda^k)$ implies $z^k_i/z^k_j = q_i/q_j$ for $A_{:,i} = A_{:,j}$, throughout exact iteration. Zero-column events retain their proposal. Proposition 3 proves the row-wise Pythagorean identity invoked in Section 3.3, $D^+_{\mathrm{KL}}(u\|z^k) = D^+_{\mathrm{KL}}(u\|z^{k+1}) + D^+_{\mathrm{KL}}(z^{k+1}\|z^k)$ for every fixed positive feasible u; it controls divergence to a fixed positive feasible point, although shared events couple relay residuals. These statements do not require preservation of every event ratio or source throughput and do not imply finite-budget success. Appendix E checks the identity, the identical-column and zero-column invariants, and the event-splitting consistency numerically on constructed problems.

With policy, time cost, proposal, and accepted menu fixed, a positive feasible alternative u_e increases the quantity term of the cost C_e of Section 3.4 by $\eta_e(u_e)/s_0$. Its unnormalized weight in Eq. (5) is therefore multiplied by $\exp\{-\eta_e(u_e)/(s_0 T)\}$. This is an interpretation of the objective, not a new reweighting operation; its consequences for constructed menus are illustrated in Appendix E. Practical selection evaluates the accepted numerical completion without subtracting an unobserved excess term.

C.6 Temporal completion

For source pattern window $[\ell_0, u_0]$, set $\Delta_0 = \max(u_0 - \ell_0, 1)$. Independently of quantity completion, temporal completion solves

$$\tau^\star_e = \operatorname*{arg\,min}_{\tau \in \mathcal{T}_{\mathcal{B}}(S_e)} D_\tau(\tau, \widehat{\tau}_e), \qquad D_\tau(\tau, \widehat{\tau}) = \frac{\sum_{i \in F_e} (\tau_i - \widehat{\tau}_i)^2}{2 \max(1, |F_e|)\Delta_0^2}, \tag{16}$$

where $\mathcal{T}_{\mathcal{B}}(S_e)$ is the set of time assignments meeting the window, pin, and ordering restrictions (Appendix A). Pattern times remain within the source window and pins remain unchanged. In DAG mode, acyclic pattern graphs additionally require each incoming pattern event to precede or coincide with each outgoing pattern event at the same node. Cyclic graphs use window and pin constraints only. Fixed core attachments keep their observed times, even outside the pattern window. The problem is convex: box-only cases use clipping, and ordering-constrained cases use a numerical quadratic solve. Quantity and time completion are separable.

C.7 Optional calibrated selection

An optional correction reweights completed candidates within protected classes $c(e)$, without changing their attributes. Write $c(e) = (\mathrm{TypedWidth}(S_e), U_e)$, where U_e records the signed multiset of added and removed endpoint pairs whose two endpoints are unanchored. Anchors come from endpoints of pinned pattern events P and positive context-flow summaries. Membership in E^{att} alone does not make an endpoint an anchor, although its fixed observations remain protected through P_c. These summaries do not insert context quantities into relay equations. Typed width is the intermediary count for scatter–gather and the pattern-event count otherwise. The six structural features ξ_e are pair uniqueness, mean $\log(1 + \mathrm{degree})$ for total, incoming, and outgoing degrees, leaf proportion, and a rewiring indicator. The correction is

$$p_\beta(e \mid c) \propto p_{\mathrm{base}}(e \mid c) \exp(t_c \beta^\top \xi_e), \qquad p_\beta(c) = p_{\mathrm{base}}(c). \tag{17}$$

The largest admissible $t_c \in [0, 1]$ is chosen so that

$$D_{\mathrm{KL}}(p_\beta(\cdot \mid c) \| p_{\mathrm{base}}(\cdot \mid c)) \le \delta_{\mathrm{KL}}. \tag{18}$$

The default budget is 0.03; singleton or constant-score classes remain unchanged. The same cap is applied during fitting, gating, and generation. Parent-balanced reconstruction loss with ridge regularization fits the correction on calibration parents. The parameterization $\beta = U\theta$ uses an orthonormal basis of identifiable within-class feature directions, with $\theta_j \in [-4, 4]$. A separate gate requires at least eight evaluable parents, target-menu/completion coverage of 0.5, and a margin above 10^{-4} after subtracting 1.96 standard errors from their mean log-loss gain. The gate is an engineering safeguard; downstream improvement remains an empirical question. Otherwise the correction is exactly the identity. Shared class and conditional draws make learned and identity-calibrated outputs identical. No downstream test score activates the correction.

C.8 Fixed-topology ablation

The `dba_kl_only` arm retains the original topology and draws one attribute proposal using the same diffusion prior, pins, unit conventions, temporal rules, and terminal DBA-KL completion. Its proposal uses the same keyed noise as the corresponding no-edit state. It invokes neither the structural policy nor categorical selection or calibration. Failure is recorded as an omission rather than replaced by an edited graph or copied source.

This defines the no-edit output, not an isolated runner timing: shared candidate completion across arms is not a per-arm cost measurement.

C.9 Stepwise and multi-round sampling variants

All comparison variants share the fitted prior, source boundary, keyed initial noise, and frozen structural candidate bank, and none edits topology during denoising. Free corrected times are re-encoded in the source pattern window with fractions clipped to $[10^{-12}, 1 - 10^{-12}]$; degenerate windows use fraction $1/2$ and clipping is recorded. Fixed observations are restored before correction and fixed encodings are restored afterward.

ECI-DBA. Our graph/EDM adaptation of extrapolate–correct–interpolate sampling (Cheng et al., 2025) takes a fresh denoiser prediction and a fresh completion in each inner round and returns to the original noise between rounds; one round per noise level is the default and three rounds the multi-round comparison. Noise is not resampled. Because the correction is DBA-KL, each call inherits the fixed-reference inner-solver properties, but a changing reference across rounds means the trajectory as a whole has no single accounting identity. An unsuccessful DBA-KL or temporal correction rejects the candidate.

Multi-round DBA-KL with projected EDM. This variant preserves the current estimated noise residual between inner rounds instead of returning to the original noise, takes a corrected-endpoint Euler step, and re-evaluates the denoiser at each of the three default rounds. It is a specified multi-round extension of projected EDM, distinct from the ECI update.

PCFM with fixed-reference log projection. The positive PCFM adaptation (Utkarsh et al., 2025) uses one-step endpoint prediction, one damped log-coordinate Gauss–Newton correction per noise level, zero interpolation penalty, and the same fixed-initial-noise interpolation; intermediate corrections need not meet the terminal relay tolerance. The terminal projector retains the last raw decoded prediction as a fixed reference q_{ref} and minimizes $\frac{1}{2}\|y - \log q_{\mathrm{ref}}\|_2^2$ subject to $A \exp(y) = b$, using rank-aware Jacobian solves, row scaling, scalar damping, and merit backtracking. Acceptance requires positive finite quantities, the physical relay test, and first-order stationarity. This is not a global nearest-point certificate for the generally nonconvex log-coordinate problem, and because the objective differs from quantity-space KL, excess KL cost does not imply failure at the log-distance objective. There is no DBA-KL fallback.

Equality-only controls. ECI-naive and PCFM-naive replace flow correction with the affine Euclidean projection $\min_z \frac{1}{2}\|z - q\|_2^2$ subject to $Az = b$, without positivity. Nonpositive values cannot

be re-encoded as log quantities and cause rejection. At their default one-round, zero-penalty settings their exact-arithmetic trajectories coincide apart from PCFM-naive's additional terminal affine projection, so they are equality-only controls rather than independent evidence from two distinct default trajectories.

Common scoring and computational budget. Every research candidate has a frozen terminal-EDM bank proposal. Common finalization restores fixed observations, performs temporal completion and the final checks, and validates the stored graph; selection scores each accepted output by its scaled KL distance to the *bank* reference plus the temporal term, using Eq. (5) without the terminal calibrator. That shared reference differs from ECI's changing references and from PCFM's own terminal reference, and only terminal DBA-KL targets the bank-reference KL minimum by construction. The accounting identity compares outputs with respect to one shared (A, b, q) even when each sampler's own correction reference differs. All methods exclude candidates without a valid bank proposal and record their own subsequent completion failures and source coverage.

For L noise levels the terminal path uses L denoiser evaluations and one flow solve per candidate, whereas multi-round projected-EDM and ECI paths use LR_D and LR_E of each; PCFM uses L intermediate corrections and one terminal projection. Multi-round methods therefore do not have the same compute budget as one-round methods, and bank construction, common finalization, and persistence add work beyond these trajectory counts.

Terminal-only completion probes. The generation-only diagnostic holds the uncorrected EDM trajectory fixed and calls either DBA-KL or the fixed-reference log projector once at its endpoint, both from the same fresh positive proposal. These are single-candidate probes, not additional full-population downstream results.

C.10 Computational scope

DBA-KL uses sparse signed-incidence arithmetic without a dense inverse. Local balancing work per sweep is proportional to the number of nonzeros, but the reference implementation also scans full vectors for numerical safety after row updates. Total work includes candidate construction, denoising, potentially many solver sweeps, temporal optimization, and validation. Neither linear end-to-end runtime nor completion within a fixed iteration budget is guaranteed; Section 4.5 measures the sweep growth on coupled relay chains and the per-call cost of the alternative solvers.

D Downstream comparison details

D.1 Result series and repetition units

The original native-comparison tables are generated by `scripts/build_rq_tables.py` from the eleven-configuration reporter's `average_AP.csv` and `summary_numeric.csv` for each dataset (reporter revision `eleven_comparison.staged_waves.v1`), copied verbatim to `results/rq/`; `results/rq/REPORT_PROVENANCE.json` records the source archive, run directories, fitted-prior identifiers, script hash, and file hashes. The original AP tables print means and SDs rounded to three decimals; the reporter's six-decimal values are in `results/rq/`. The reporter aggregates the saved final test metrics directly; no earlier three-seed summary is pooled into these tables.

LI-Small and HI-Small each record 220 complete fits on the balanced four-model, eleven-configuration grid (detector seeds 0–4, five per cell), and LI-Medium records 176 (seeds 0–3, four per cell); no fit is missing or non-evaluable. FraudGT is outside the LI-Medium model scope and enters no average on any dataset.

The external-generator rows (HARPOON, GDSS) come from a separate run set with the same seed set. We take down the per-evaluator AP mean and SD at three decimals for each dataset and the rows are placed after the saved series in Tables 1, 5, and 6–8.

The reporter filters the comparison series (the ECI, PCFM, and multi-round arms) to generation seed 0 and one compatible fitted-prior identifier; the Real, Learned, Calibrated, and fixed-topology series is its separately saved native population. For a complete model/seed grid, the reporter first forms the

Table 4: Documented inherited detector settings. These are the conventions of the inherited trainer, not verified per-run configurations: the saved exports do not carry per-run overrides, and no matched-comparison claim is made from them.

Field	Documented setting / verification required
Architecture	Two graph layers and an edge-classification head.
Hidden width	64 in the inherited trainer; verify each selected run.
Optimization	AdamW; 150 optimizer updates; initial learning rate 0.003.
Regularization	Weight decay 0.001; gradient clipping at 2.
Learning-rate schedule	Cosine decay to 10^{-5} in the inherited configuration.
Positive BCE weight	$\min(40, N_-/N_+)$ using the arm's actual target counts.
Checkpoint	Final epoch; no best-test-AP checkpoint selection.
Thresholded metrics	Probability threshold 0.5; separate from AP.
Message policy	`full` (all events in the training graph) in every series; see Appendix D.3.

four-model average $a_{c,s}$ and then computes

$$s_c = \sqrt{\frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}}(a_{c,s} - \bar{a}_c)^2}.$$

Relative differences in Section 4 are calculated from the reporter's six-decimal means, so they may differ in the last digit from values recomputed from the printed three-decimal entries. Paired uncertainty requires the original seedwise differences; no p-values or confidence intervals are inferred here.

The additional no-flow-completion results are kept separately in `results/rq2_no_flow/`; Table 2 prints their supplied six-decimal means and SDs. The reported relay audit and its scope are in Appendix D.5.

D.2 Shared downstream protocol and fields to verify

Let E^- be the original normal background, with chronological negative targets E^-_{tr} and E^-_{te} in the 80/20 split. The positive case lists define disjoint E^+_{tr} and E^+_{te}. The intended graph and supervision sets are

$$\begin{aligned} G^{\mathrm{real}}_{\mathrm{tr}} &= E^- \cup E^+_{\mathrm{tr}}, & Y^{\mathrm{real}}_{\mathrm{tr}} &= E^-_{\mathrm{tr}} \cup E^+_{\mathrm{tr}}, \\ G^{c}_{\mathrm{tr}} &= E^- \cup \widetilde{E}^+_{\mathrm{tr},c}, & Y^{c}_{\mathrm{tr}} &= E^-_{\mathrm{tr}} \cup \widetilde{E}^+_{\mathrm{tr},c}, \\ G_{\mathrm{te}} &= E^- \cup E^+_{\mathrm{te}}, & Y_{\mathrm{te}} &= E^-_{\mathrm{te}} \cup E^+_{\mathrm{te}}. \end{aligned}$$

Training loss is computed only on training targets. The graph shown here is the data population; every series was trained with the same message-passing policy (Appendix D.3). Counts of synthetic positives and retained cases can differ across arms and are not silently equalized. A common-source/count control is a separate experiment, not a change to these existing scores.

The documented adapter uses standardized log quantity, standardized log sender gap, and log directed-pair occurrence index as edge channels, with four incident-event aggregates and an identity-derived node scalar. Scaling is fitted on original training targets. The inherited synthetic adapter reconstructs gap features from original training-positive gaps, so direct timestamp fidelity belongs to RQ3, not to an interpretation of AP alone. Per-run adapter, model-source, prepared-population, and test-row identities are not recoverable from the saved exports, so the detector settings listed below are the documented conventions of the inherited trainer rather than verified per-run configurations.

The LI-Medium run uses a hidden width of 32 for 150 epochs, and a cosine floor of 10^{-4} for that dataset, which differ from the inherited LI-Small/HI-Small settings above. Absolute AP is therefore not comparable across datasets; every comparison in Section 4 is within one dataset.

D.3 Comparability across result series and sampler revisions

All detector runs in all series were trained with the same message-passing policy, `full`, in which every event of the training graph carries messages. The reporter nevertheless flags a `message_policy` difference on all three datasets, because the run metadata of the Real, Learned,

Table 5: Four-GNN average AP (mean ± population SD over detector seeds) for the ten saved configurations of Table 1 and the two external generators. Per-evaluator means and SDs are in Tables 6–8.

Configuration	LI-Small	HI-Small	LI-Medium
Real (TRTR)	0.207 ± 0.014	0.397 ± 0.015	0.061 ± 0.031
Learned DBA-KL (terminal)	0.193 ± 0.016	0.406 ± 0.010	0.050 ± 0.028
Calibrated DBA-KL	0.193 ± 0.016	0.404 ± 0.008	0.050 ± 0.028
Fixed-topology DBA-KL	0.192 ± 0.018	0.398 ± 0.010	0.053 ± 0.029
DBA-KL/projected EDM (3 rounds)	0.194 ± 0.016	0.404 ± 0.008	0.050 ± 0.028
ECI-DBA (1 round)	0.184 ± 0.014	0.391 ± 0.007	0.045 ± 0.024
ECI-DBA (3 rounds)	0.180 ± 0.011	0.382 ± 0.009	0.041 ± 0.021
PCFM-log	0.189 ± 0.014	0.396 ± 0.011	0.045 ± 0.024
ECI-naive (equality only)	0.182 ± 0.017	0.362 ± 0.014	0.036 ± 0.019
PCFM-naive (equality only)	0.182 ± 0.017	0.362 ± 0.014	0.036 ± 0.019
HARPOON (raw output)	0.108 ± 0.033	0.249 ± 0.036	0.038 ± 0.003
GDSS (raw output)	0.072 ± 0.024	0.245 ± 0.036	0.032 ± 0.003

Calibrated, and fixed-topology series still carries the label `legacy-ddp` from an earlier configuration template that was not updated when those runs were launched; the training code path was `full` for every run. The flag is therefore a stale metadata label, not a protocol difference, and the exported files are kept as emitted rather than edited to silence it. All cross-series comparisons in Section 4 use one detector protocol.

The earlier HI-Small PCFM series is explicitly loaded from `sampler_continuation` and labeled log-root; it is not relabeled as fixed-reference, and the fixed-reference PCFM-log series supplies the log-distance terminal-projection comparison used in the main text. On LI-Small and LI-Medium the earlier series repeats the fixed-reference algorithm under a separate experiment identity. The earlier and current ECI-DBA and PCFM entries remain distinct saved series even when rounded scores coincide. Eleven columns represent saved configurations, not eleven independent algorithm families.

D.4 Four-GNN average and per-evaluator AP

Table 5 gives the four-GNN average and its population SD across detector seeds for the ten saved configurations of Table 1 and the two external generators; this is the statistic used in Section 4.2. Tables 6, 7, and 8 give the per-evaluator means and SDs behind it for LI-Small, HI-Small, and LI-Medium: the 132 saved per-evaluator means (eleven configurations, including the earlier saved PCFM series of Appendix D.3) and the 24 exported external means. On every dataset, every external cell is below the lowest DBA-KL cell and the lowest adaptation cell of its evaluator. No FraudGT score enters any four-model average, and smart-grid values are not inferred from these measurements.

D.5 No-flow-completion ablation: predictive utility and relay invalidity

Population and result provenance. Table 2 contains only the additional `rq2_no_flow` arm on the four common evaluators. The reported configuration uses one frozen native generator per dataset and generation draw 0; detector repetitions follow the reference grids (five seeds for LI-Small and HI-Small, four for LI-Medium). The means and SDs are retained as supplied, including the SD of each seedwise four-GNN average. The tabulated six-decimal model means can differ from the supplied aggregate in the final digit because of rounding. The normalized input transcriptions, comparator means, and report provenance are retained in `results/rq2_no_flow/`. This manuscript integrates supplied completed run summaries and audit reports; we ran no further generation or detector experiment. The ablation's per-evaluator AP is Table 2 in the main text.

What is removed. The ablation retains structural proposals, learned structural preference, diffusion-proposed positive quantities, pinned values, and temporal completion. It omits every quantity-correction call, flow-feasibility filter, and flow-distortion term in candidate selection. It can

Table 6: LI-Small: all eleven saved configurations on the four retained evaluators and their equal-weight average. Entries are AP means rounded to three decimals with the population SD below in parentheses (for the average, the SD of the seedwise four-GNN average); 5 detector seeds per cell (220 complete fits) in the saved series. The earlier saved PCFM series is a repeat of the fixed-reference algorithm under a separate experiment identity (Appendix D.3).

Configuration	EdgeSAGE	EdgeGCN	EdgeGAT	EdgeGIN	Average
Real (TRTR)	0.333 (0.061)	0.200 (0.032)	0.057 (0.006)	0.241 (0.024)	0.207 (0.014)
Learned DBA-KL (terminal)	0.329 (0.074)	0.182 (0.036)	0.062 (0.010)	0.201 (0.029)	0.193 (0.016)
Calibrated DBA-KL	0.329 (0.074)	0.182 (0.036)	0.062 (0.010)	0.201 (0.029)	0.193 (0.016)
Fixed-topology DBA-KL	0.330 (0.074)	0.182 (0.034)	0.059 (0.010)	0.197 (0.028)	0.192 (0.018)
DBA-KL/projected EDM (3 rounds)	0.334 (0.076)	0.181 (0.033)	0.060 (0.008)	0.202 (0.028)	0.194 (0.016)
ECI-DBA (1 round)	0.316 (0.074)	0.173 (0.039)	0.053 (0.009)	0.194 (0.027)	0.184 (0.014)
ECI-DBA (3 rounds)	0.309 (0.073)	0.168 (0.038)	0.050 (0.007)	0.194 (0.028)	0.180 (0.011)
PCFM-log	0.317 (0.075)	0.178 (0.037)	0.056 (0.009)	0.203 (0.031)	0.189 (0.014)
PCFM-log (earlier saved series)	0.317 (0.075)	0.178 (0.037)	0.055 (0.008)	0.203 (0.031)	0.188 (0.014)
ECI-naive (equality only)	0.304 (0.082)	0.176 (0.033)	0.053 (0.005)	0.194 (0.040)	0.182 (0.017)
PCFM-naive (equality only)	0.304 (0.082)	0.176 (0.033)	0.053 (0.005)	0.194 (0.040)	0.182 (0.017)
HARPOON (raw output)	0.213 (0.034)	0.088 (0.040)	0.027 (0.010)	0.104 (0.032)	0.108 (0.033)
GDSS (raw output)	0.152 (0.033)	0.055 (0.024)	0.019 (0.009)	0.064 (0.020)	0.072 (0.024)

therefore select different topologies or retain different sources from terminal Learned. It preserves original amounts only incidentally, so it is not a topology-only control, and its no-edit candidate remains available. The higher aggregate AP relative to Real supports useful topology-enabled synthesis. It speaks to the pipeline as a whole: it shows neither that every evaluator improves nor that editing alone causes the difference. Relative to terminal Learned, the aggregate increases are 8.73%, 2.87%, and 23.11%, so the two arms differ in utility as well as in validity.

Audit protocol. The supplied reports describe fresh checks of the actual selected graph bytes, authenticated against their export hashes before deserialization, using the original native relay checker. The modeled event set consists of core transactions, including fixed core embeds, and excludes halo/context events. Relays are reconstructed from each candidate's modeled incidence; pure sources and sinks have no equation. With the original frozen unit conversion, each relay a passes precisely when

$$|r_a| \le 10^{-10} + 10^{-8}\max(I_a, O_a), \qquad \rho_a = \frac{|r_a|}{10^{-10} + 10^{-8}\max(I_a, O_a)} \le 1,$$

where r_a is the compensated signed sum of incoming and outgoing modeled quantities. A relay-containing graph passes only if every relay passes. Relay-free outputs are N/A and excluded from this failure-rate denominator; omissions remain separate. The reported residual quantiles summarize individual relays within each dataset, not graphs or independent generator runs.

Observed conservation failure. Across 1162 offered sources, 1095 selected graphs are emitted and 67 sources are omitted. Every audited relay-containing graph fails: $88/88$ on LI-Small, $250/250$ on HI-Small, and $272/272$ on LI-Medium, totaling $610/610$. At the individual-relay level, 1916 of 1918 checks fail; the two passing relay checks do not make any relay-containing graph fully balanced. All 1095 emitted graphs have positive core quantities, with no empty-core graph, unresolved audit error, or disagreement with the saved flow audit. The 485 relay-free graphs carry no conservation equation, so they say nothing about balancing.

Table 7: HI-Small: all eleven saved configurations on the four retained evaluators and their equal-weight average. Entries are AP means rounded to three decimals with the population SD below in parentheses (for the average, the SD of the seedwise four-GNN average); 5 detector seeds per cell (220 complete fits) in the saved series. The earlier saved PCFM series uses the log-root revision; PCFM-log is the fixed-reference revision (Appendix D.3).

Configuration	EdgeSAGE	EdgeGCN	EdgeGAT	EdgeGIN	Average
Real (TRTR)	0.502 (0.032)	0.393 (0.024)	0.209 (0.023)	0.484 (0.022)	0.397 (0.015)
Learned DBA-KL (terminal)	0.538 (0.022)	0.385 (0.015)	0.214 (0.021)	0.488 (0.025)	0.406 (0.010)
Calibrated DBA-KL	0.538 (0.022)	0.385 (0.015)	0.205 (0.025)	0.488 (0.025)	0.404 (0.008)
Fixed-topology DBA-KL	0.523 (0.024)	0.388 (0.021)	0.192 (0.025)	0.489 (0.009)	0.398 (0.010)
DBA-KL/projected EDM (3 rounds)	0.545 (0.025)	0.387 (0.020)	0.199 (0.034)	0.484 (0.027)	0.404 (0.008)
ECI-DBA (1 round)	0.526 (0.019)	0.374 (0.029)	0.189 (0.031)	0.473 (0.040)	0.391 (0.007)
ECI-DBA (3 rounds)	0.518 (0.020)	0.369 (0.029)	0.174 (0.026)	0.468 (0.020)	0.382 (0.009)
PCFM-log	0.525 (0.021)	0.374 (0.030)	0.199 (0.031)	0.485 (0.022)	0.396 (0.011)
PCFM (log-root, earlier saved series)	0.525 (0.021)	0.374 (0.030)	0.198 (0.030)	0.485 (0.022)	0.395 (0.010)
ECI-naive (equality only)	0.517 (0.020)	0.338 (0.026)	0.146 (0.039)	0.446 (0.034)	0.362 (0.014)
PCFM-naive (equality only)	0.517 (0.020)	0.338 (0.026)	0.145 (0.038)	0.446 (0.034)	0.362 (0.014)
HARPOON (raw output)	0.365 (0.067)	0.226 (0.034)	0.169 (0.021)	0.236 (0.035)	0.249 (0.036)
GDSS (raw output)	0.379 (0.047)	0.221 (0.036)	0.163 (0.023)	0.217 (0.049)	0.245 (0.036)

Interpretation and scope. These 610 relay-containing mutants are invalid under our skeleton-local, lossless, zero-net-accumulation model: violating one required relay equation suffices, regardless of their downstream AP. The result shows that positive generated quantities and retained predictive utility can coexist with invalid scenarios, which motivates explicit flow enforcement when validity matters. DBA-KL is one such mechanism among others, and conservation need not improve AP. No-flow removal changes quantities, eligibility, source retention, and selection together, so the cause of the AP difference stays unidentified. The audit also covers relay balance alone: pins, vocabulary, timestamps, task labels, and topology diversity come from the separate RQ3 audit.

D.6 Saved-skeleton audit and external quantity validity (RQ3)

The audit reads the saved generation banks (generation draw 0 of each dataset's recorded fitted model) without regeneration or numerical repair; it is saved-population evidence, not a new generation or detector run. For each dataset and configuration it covers one stored draw per offered source, so across-configuration differences are not across-draw diversity, and no generator-seed uncertainty is estimated. Relays are reconstructed from the modeled core incidence, including fixed core attachments and excluding context-only events; each relay's residual is normalized by the recorded absolute-plus-relative tolerance, and relay-free graphs are excluded from the relay-pass denominator but retained in every other denominator. A graph is valid when the schema, pins and context, positive core quantities, relay balance, time-order, source-family, and closed observed-account-vocabulary predicates all pass; unknown checks are exposed, not counted as passes. Fixed-account incidence equality compares complete canonical multisets of typed events with editable event identifiers and numerical attributes ignored; the supplementary anonymous and pin-aware isomorphism preserves multiplicity and currency. Original records are descriptive references and are not required to satisfy the zero-relay scenario model, which is a construction rule for generated scenarios. The audit supports quantitative validity, retention, source-relative structural variation, and quantity and time discrepancies; it does not validate task intent, production robustness, or coverage of every feasible topology. The fixed-topology and terminal Learned comparison is summarized in Table 3 in the main text.

Table 8: LI-Medium: all eleven saved configurations on the four retained evaluators and their equal-weight average. Entries are AP means rounded to three decimals with the population SD below in parentheses (for the average, the SD of the seedwise four-GNN average); 4 detector seeds per cell (176 complete fits) in the saved series. FraudGT is outside the LI-Medium scope. The earlier saved PCFM series is a repeat of the fixed-reference algorithm under a separate experiment identity (Appendix D.3).

Configuration	EdgeSAGE	EdgeGCN	EdgeGAT	EdgeGIN	Average
Real (TRTR)	0.110 (0.084)	0.018 (0.003)	0.032 (0.011)	0.084 (0.058)	0.061 (0.031)
Learned DBA-KL (terminal)	0.090 (0.076)	0.016 (0.003)	0.029 (0.006)	0.067 (0.051)	0.050 (0.028)
Calibrated DBA-KL	0.090 (0.076)	0.016 (0.003)	0.029 (0.006)	0.067 (0.051)	0.050 (0.028)
Fixed-topology DBA-KL	0.093 (0.078)	0.016 (0.003)	0.030 (0.007)	0.070 (0.052)	0.053 (0.029)
DBA-KL/projected EDM (3 rounds)	0.090 (0.076)	0.016 (0.002)	0.029 (0.006)	0.067 (0.052)	0.050 (0.028)
ECI-DBA (1 round)	0.080 (0.068)	0.015 (0.002)	0.025 (0.004)	0.059 (0.045)	0.045 (0.024)
ECI-DBA (3 rounds)	0.073 (0.062)	0.014 (0.002)	0.024 (0.003)	0.053 (0.038)	0.041 (0.021)
PCFM-log	0.079 (0.067)	0.015 (0.002)	0.026 (0.004)	0.060 (0.045)	0.045 (0.024)
PCFM-log (earlier saved series)	0.079 (0.067)	0.015 (0.002)	0.026 (0.004)	0.060 (0.045)	0.045 (0.024)
ECI-naive (equality only)	0.060 (0.051)	0.013 (0.002)	0.024 (0.002)	0.047 (0.038)	0.036 (0.019)
PCFM-naive (equality only)	0.060 (0.051)	0.013 (0.002)	0.024 (0.002)	0.047 (0.038)	0.036 (0.019)
HARPOON (raw output)	0.069 (0.026)	0.011 (0.006)	0.022 (0.006)	0.048 (0.004)	0.038 (0.003)
GDSS (raw output)	0.058 (0.018)	0.008 (0.004)	0.019 (0.005)	0.043 (0.001)	0.032 (0.003)

Table 9: Reported saved-graph relay audit of `rq2_no_flow`, generation draw 0. Graph failure rates use only relay-containing outputs; relay-free graphs remain N/A. Normalized residuals use the recorded 10^{-10} absolute and 10^{-8} relative tolerances.

Measurement	LI-Small	HI-Small	LI-Medium
Offered sources	159	459	544
Emitted selected graphs	153	431	511
Omitted sources	6	28	33
Graphs containing relays	88	250	272
Relay-free graphs (N/A)	65	181	239
Relay-containing graphs passing	0	0	0
Relay-containing graphs failing	88	250	272
Relay-graph failure rate	100.00%	100.00%	100.00%
Failed / audited relays	260/260	763/764	893/894
Median normalized residual	3.49815×10^7	3.16712×10^7	3.88214×10^7
P95 normalized residual	9.14988×10^7	9.51063×10^7	9.76835×10^7
Maximum normalized residual	9.99954×10^7	9.99949×10^7	9.99975×10^7
Core-positive emitted graphs	153	431	511
Empty-core graphs	0	0	0
Unresolved graph-audit errors	0	0	0
Disagreements with saved audit	0	0	0

Per-configuration counts for all ten flow-corrected configurations are in the supplied `results/rq3/audit/` exports. Within this population, every invalid generated graph fails only the vocabulary predicate (schema, positivity, pins and context, relay balance, time order, and family have zero failures in each of these flow-corrected arms); the vocabulary failures are edited graphs and concentrate in the largest family, STACK (69–79% of them). The adaptations share the structural pipeline and stay within eight valid graphs of Learned on every dataset. Omissions are dominated by sources outside the grammar's support (5, 26, and 33) and, for fixed topology, by pins that are infeasible under

the zero-relay model (3, 15, and 12) or non-converged solves (4, 13, and 18), so editing recovers some sources that fixed topology cannot emit. Where the stored candidate diagnostics are complete (the comparison series), the positivity-aware completions certify 88.9–93.8% of candidates and the equality-only controls 82.5–88.4%, consistent with their lower emission. Across the ten generated configurations, 45%, 52%, and 56% of the sources receive at least two distinct valid fixed-account structures, a conditional observation that completion cost enters selection, not draw diversity. On common valid sources, editing raises the source-window-normalized time $\mathcal{W}_1$ from 0.08 to 0.21–0.23 and changes the log-quantity $\mathcal{W}_1$ by $+0.20$, $+0.04$, and -0.04 relative to fixed topology. The audit's summary, per-family, candidate-acceptance, omission, common-source, and within-source files are in `results/rq3/audit/` (the per-graph and per-relay files are identified by hash); `scripts/build_rq3_tables.py` generates the tables from them.

The external generators' raw outputs are not audited structurally: their downstream utility already places them below every DBA-KL configuration and every adaptation (Section 4.2), so we report only their quantity conservation on the same three datasets. The mean absolute relay-balance error of HARPOON's raw outputs is 4.2×10^7, 2.6×10^8, and 4.2×10^9 common units on LI-Small, HI-Small, and LI-Medium, and GDSS's 2.1×10^8, 4.3×10^9, and 8.0×10^9; 0.35–2.32% of their generated quantities are negative, and their Wasserstein-1 distances to the original quantity distribution are 1.5×10^7 to 1.3×10^8 (HARPOON) and 4.6×10^7 to 2.8×10^9 (GDSS).

Every DBA-KL output has zero relay-balance error within tolerance and strictly positive quantities by construction. The metrics are kept in `results/rq3/external_quantity_validity.csv`.

D.7 Completion cost study (RQ4)

Problems. For each structure and relay count r we build a signed incidence over events with a planted positive balanced witness z: *chain*, r relays in series over $r+1$ events; *cycle*, r relays in a ring over r events ($r \geq 2$); *independent*, r decoupled relays with one incoming and one outgoing event each; *fan*, r decoupled relays with three incoming events drawn uniformly from $[0.5, 2]$ and two outgoing events splitting their sum by a uniform Dirichlet draw; *layered*, a width-two stack with r relays in $\lceil r/2 \rceil$ layers between two source and two sink accounts, every relay of a layer sending to every relay of the next, whose witness sums one random source-to-sink path of weight uniform in $[0.5, 2]$ through every event. Chain and cycle witnesses are unit flows; independent witnesses are uniform in $[0.5, 2]$. Then $\lfloor 0.25\, n \rfloor$ of the n events (at most $n-1$), chosen uniformly, are pinned at their witness values, which determine b and the fixed inflow and outflow of every relay; the proposal on the free events is $q = z \odot \exp(\sigma\epsilon)$ with standard normal ϵ and $\sigma \in \{0.5, 1.5\}$; the source scale is the total witness flow. Each cell has 50 instances (seed 20260926, instance i at seed $20260926000 + i$), and every method receives the same instance. The 38 cells (five structures, $r \in \{1, 2, 4, 8\}$, two values of σ; no one-relay cycle) give 7,600 timed calls.

Methods, timing, and reference. Native DBA-KL runs the production solver (tolerance 10^{-8}, balance tolerance 10^{-10}, at most 2000 sweeps, fixed inflow and outflow passed explicitly). The same-KL Newton solve is the rank-reduced dual Newton reference of Appendix E (relative tolerance 10^{-11}, absolute 10^{-12}, at most 500 iterations, stationarity and duality-gap checks). The PCFM-log projection uses q as both reference and initial point with at most 2000 iterations, and the positive QP is the supplied SLSQP implementation with its KKT check and floor 10^{-12}; the four implementations are those of Appendix E. Acceptance is each solver's own certificate on the original constraints. Every call is single-threaded (OpenMP, MKL, and OpenBLAS threads set to one) and sequential on an x86-64 machine with two CPUs, Python 3.11 and NumPy 2.4, after warm-up calls of every method; each instance and method is timed three times and the median is kept, and P50 and P95 are over the 50 instances, with failed calls kept in the timing. The Newton solution is computed once more, untimed, as the reference, and the excess KL from every method's output to it, per unit of source scale, records solution agreement (Table 10). The study took 191 s in total.

Reading the table. On the chain, cycle, and independent structures every relay has one incoming and one outgoing event, so relay balance equates event quantities and both the KL projection and the log-coordinate projection return the geometric mean of each equated class (a pinned member fixes the class); there PCFM-log coincides with the KL optimum (excess below 4×10^{-14}), whereas

Table 10: Per-call completion cost, P50 wall time in ms (single thread, median of three timed calls, 50 instances per cell, log-noise SD 0.5). DBA-KL columns give the median number of sweeps in parentheses; the last three columns give the same-KL Newton solve, the fixed-reference PCFM-log projection, and the positive QP on the same eight-relay problems. P95, log-noise SD 1.5, iteration counts, acceptance, and excess KL are in the supplied `results/rq4_cost/` exports; the full per-cell table is not compiled into this paper.

	DBA-KL, P50 ms (median sweeps)				$r = 8$, P50 ms		
Structure	$r = 1$	$r = 2$	$r = 4$	$r = 8$	Newton	PCFM-log	QP
Chain	0.41 (1)	1.56 (14)	8.32 (60)	24.3 (106)	2.25	15.4	4.17
Cycle	–	0.44 (1)	2.17 (14)	8.89 (36)	2.16	14.5	4.11
Independent	0.42 (1)	0.44 (1)	0.50 (1)	0.60 (1)	2.19	15.3	4.42
Fan (3 in, 2 out)	0.41 (1)	0.44 (1)	0.51 (1)	0.62 (1)	2.13	16.7	4.63
Layered (width 2)	0.41 (1)	0.44 (1)	2.47 (15)	11.6 (47)	2.47	17.9	4.53

on the fan and layered structures the two objectives differ (median excess 7×10^{-4} to 6×10^{-2}, maximum 0.57). Where the pins fix every class (chains and cycles with four or more relays at this pin fraction) the feasible set is a single point, the QP's excess is zero as well, and the cell measures only the cost of reaching a known point; DBA-KL still needs 60–113 median sweeps on these chains because each sweep moves information one relay along the chain. The QP time is nearly constant across cells (4.0–4.8 ms with one or two iterations), which reflects fixed solver overhead rather than problem size. The relay counts of the audited banks come from the audit's per-graph file (`results/rq3/audit/relay_count_distribution.csv`): the shares of sources with at most 1, 2, 4, and 8 relays are 80%, 86%, 90%, and 93% on LI-Small, 74%, 83%, 89%, and 95% on HI-Small, and 73%, 82%, 88%, and 94% on LI-Medium, with median 1 and maxima 15, 16, and 16; among emitted terminal Learned graphs, 87–88% have at most two relays and 97–98% at most eight. The study measures the per-call cost of the reference implementations on constructed problems only: no sampler, denoiser, candidate construction, temporal completion, validation, or AML data is involved, and no runtime is claimed for the generator. Data, harness, and table script are in `results/rq4_cost/` (`per_call.csv`, `summary.csv`, `META.json`), `scripts/rq4_cost/run_cost_study.py`, and `scripts/rq4_cost/build_rq4_cost_tables.py`.

E Numerical verification of the distortion accounting

Summary. This appendix checks the identity of Eq. (3), the accuracy of the DBA-KL routine, and the selection law of Eq. (5) on constructed problems whose optimum is known before any solver runs. It is a verification study: the identity is a theorem, the excess of a completion that optimizes another objective is nonzero by definition, and the selection effects below are properties of the designed inputs. It establishes neither the prevalence of excess distortion on the AML data nor any downstream or runtime effect. On 120 constructed signed-incidence problems (20 per family: split relays, parallel chains, diamonds, cycles, layered DAGs, and relays with an unconstrained event), a positive feasible $z^\star$ is constructed first and the proposal is frozen as $q = z^\star \odot \exp(A^\top \lambda)$, so $z^\star$ is the KL optimum by the KKT condition. With $m = D^+_{\mathrm{KL}}(z^\star \| q)$ and $\eta = D^+_{\mathrm{KL}}(u \| z^\star)$, the 110 cases that admit a nontrivial feasible perturbation u have maximum scaled identity discrepancy $|D^+_{\mathrm{KL}}(u\|q) - m - \eta| / \max(1, |D^+_{\mathrm{KL}}(u\|q)|, m, \eta) = 3.07 \times 10^{-16}$ at the known optimum; substituting the numerical DBA-KL output for $z^\star$ raises it to 2.14×10^{-8} at native tolerances and 1.93×10^{-12} at the tight diagnostic tolerances, so finite-precision accounting deserves an explicit check; balance alone is not enough. Table 11 lists the accepted runs and the excess KL of each completion: the independent same-KL Newton solver attains the same reference, whereas PCFM-log and the positive QP, which optimize other objectives, incur median excess 2.45×10^{-3} and 7.35×10^{-3} per source scale with maxima 0.82 and 91.2. Twenty-four matched pairs with prescribed minimum/excess fractions $0.8/0.2$ versus $0.2/0.8$ are recovered in 48/48 cases (fraction error 1.04×10^{-14}). On 12 constructed three-topology menus at four temperatures, a heterogeneous excess of $0.12 s_0$ on one candidate changes the selection mode in 48/48 comparisons and a constant excess cancels, whereas the actual PCFM-log and QP completions shift the selection probabilities (median total variation 0.0219 and 0.1689) without changing the mode in any of the 48 comparisons.

Table 11: Constructed-problem completions: accepted runs and the median and maximum of $D^+_{\mathrm{KL}}(x\|z^\star)/s_0$ over accepted outputs, with the same source scale s_0 for all methods on a case. Values below 10^{-14} are at the reporting resolution. PCFM-log's four unresolved optimality stops are retained and are not labeled infeasible. Excess is measured from the KL optimum, so a nonzero value for PCFM-log or the QP reflects a different objective, not a failure on its own terms.

Completion	Accepted	Median excess$/s_0$	Max excess$/s_0$
DBA-KL (native tolerance)	120/120	$< 10^{-14}$	$< 10^{-14}$
DBA-KL (tight diagnostic)	120/120	$< 10^{-14}$	$< 10^{-14}$
Independent same-KL Newton	120/120	$< 10^{-14}$	$< 10^{-14}$
PCFM-log fixed reference	116/120	2.45×10^{-3}	0.82
Positive quadratic program	120/120	7.35×10^{-3}	91.2

Scope and reproducibility. All quantities in this section come from the included executed toy suite, not the AML runs. The primary grid contains 120 problems, 20 per family, generated with seed 20260925. There are 1–8 relay rows; quantity scales cycle through $10^{-2}, 1, 10^2, 10^4$, and the maximum construction tilt cycles through $0.25, 1, 2, 3$. This is a designed grid, not a random sample from the application distribution. Each method receives the same frozen (A, b, q). Temporal completion, structural-policy fitting, generator denoising, classifier training, and empirical speed comparisons are outside this study.

A solver-independent optimum. We form a full modeled incidence matrix C from explicit event endpoints. Positive source-to-sink path flows or positive cycle flows give $Cv^0 = 0$. For the pinned set P and free set F, set

$$A = C_{:,F}, \qquad b = -C_{:,P}v^0_P, \qquad z^\star = v^0_F, \qquad q = z^\star \odot \exp(A^\top \lambda).$$

The case normalization is $s_0 = \sum_i v^0_i$. Pins are eliminated, not optimized; no background quantity enters b. Split and chain cases contain parallel columns. Cycle rows are intentionally rank-deficient; the zero-column family includes a source-to-sink event that has no relay coefficient. Diamonds and layered DAGs supply coupled constraints. 10 cases have no free feasible direction and are excluded only from the nontrivial-perturbation test, not from solver evaluation.

For $f(z) = D^+_{\mathrm{KL}}(z\|q)$, the constructed point satisfies $Az^\star = b$ and $\nabla f(z^\star) = -A^\top\lambda$, the optimality condition of Section 3.3. Since $\nabla^2 f(z) = \mathrm{diag}(1/z_i) \succ 0$ on the positive orthant, these conditions establish its unique global optimality for the affine problem. Thus the reference is known before calling DBA-KL or any numerical competitor; it is not a DBA-KL output relabeled as ground truth. Saved endpoints, full/pinned/free quantities, proposals and multipliers allow independent reconstruction of both feasibility and stationarity.

Numerical implementations. The shipped `kl_project` and tolerance-validator function bodies are verbatim extracts from the supplied engine, identified by source hashes. Native DBA-KL uses at most 2,000 sweeps and the per-relay condition $|I_a - O_a| \le 10^{-10} + 10^{-8}\max(I_a, O_a)$; the tight diagnostic changes only these tolerances, to 10^{-13} and 10^{-12}, and not the production settings. The independent same-KL control is the existing rank-reduced global Newton dual solver with original-equation, stationarity, and dual-gap checks, and never calls DBA-KL. The PCFM adapter minimizes $\frac{1}{2}\|\log z - \log q\|_2^2$ subject to $Az = b$ from q; the positive QP minimizes $\frac{1}{2}\|z - q\|_2^2$ with a physical lower bound 10^{-12}. A feasible iterate is not automatically counted as a converged PCFM optimum.

All 120 native DBA-KL, 120 tight DBA-KL, 120 same-KL Newton, and 120 QP runs were accepted; PCFM returned 120 feasible vectors but only 116 met its full stopping criteria, and the four remaining records are retained with their reasons. Maximum relative coordinate error against the constructed optimum was 1.07×10^{-7} (native), 4.13×10^{-11} (tight), and 2.14×10^{-11} (independent). No output was clipped, substituted, or silently removed.

Identity and assumption controls. For each case with a nonempty null space we construct $u = z^\star + \alpha d > 0$ with $Ad = 0$; the 110 alternatives have maximum scaled identity discrepancy 3.07×10^{-16}. A separate 60-digit `Decimal` reader recomputes the KL values from the saved floating-point inputs and verifies the original equations, which improves arithmetic checking but is not an exact-arithmetic proof of those inputs. For any positive reference r the exact three-point expression

$D^+_{\mathrm{KL}}(u\|q) - D^+_{\mathrm{KL}}(r\|q) - D^+_{\mathrm{KL}}(u\|r) = (u-r)^\top \log(r/q)$ vanishes at the optimum for feasible u but need not vanish for an arbitrary feasible reference: for $q = (1, 4)$ and $z_1 + z_2 = 10$ the optimum is $(2, 8)$, and the feasible nonoptimal $r = (5, 5)$ with $u = (8, 2)$ leaves a discrepancy 4.158883 despite zero reference imbalance. Negative controls reject inconsistent fully fixed equations and $z_1 + z_2 = 0$ under strict positivity; no-relay controls retain q and are reported not applicable.

Allocation and matched-total checks. Identical nonzero incidence columns occur in 80 cases, where tight DBA-KL has maximum log-ratio drift 1.07×10^{-14}; the 20 zero-column cases change unconstrained quantities by at most 3.50×10^{-16} relatively; and in 120 event-splitting tests, aggregating a $0.37/0.63$ split changes the completion by at most 9.04×10^{-15} relatively, consistent with the KL chain rule. For 24 matched pairs (seed 7919) with prescribed minimum fractions 0.8 versus 0.2, the maximum relative total mismatch is 1.64×10^{-14}, the fraction error 1.04×10^{-14}, and the scaled identity error using the tight numerical reference 9.91×10^{-16}. This deliberately constructed split demonstrates the identity; real-world fault-classification accuracy is a separate question.

Controlled selection. Twelve menus (seed 4049) share one fixed source-to-relay event with two, three, or four free relay-to-sink events and normalized minimum costs $(0.01, 0.03, 0.06)$, with zero structural and time terms. At temperatures $T \in \{0.025, 0.05, 0.1, 0.2\}$ we compare $p_m \propto \exp\{-m_e/(s_0 T)\}$ with $p_{\mathrm{tot}} \propto \exp\{-(m_e + \eta_e)/(s_0 T)\}$, the selection law of Eq. (5) with zero structural and time terms. Assigning excess $0.12 s_0$ to one candidate changes the mode in 48/48 comparisons, whereas a constant $0.04 s_0$ cancels; actual PCFM and QP outputs shift the probabilities (median total variation 0.0219 and 0.1689) without changing the mode in any of the 48 comparisons. These are repeated temperatures on constructed menus, not independent AML observations, and no AP improvement or generally superior structural selection is inferred.

F Practitioner motivation and structural-scenario evaluation

F.1 Evidence boundary of the industrial motivation

The industrial motivation is a qualitative practitioner report of an evaluation bottleneck in an operational financial-risk setting: repeatedly varying amounts and timestamps on known interaction skeletons leaves decisions about different accounts and subgraphs insufficiently exercised. The reported use case includes authorized red/blue-team testing of LLM-assisted graph analysis. The organization and individual identities are omitted in this anonymous draft. No production logs, incident cohort, structured interview study, deployment measurement, or company endorsement is offered as evidence here. The report motivates a research requirement; it is not a measured prevalence estimate of shortcut behavior. The benchmarks in Section 4 remain the separately identified original datasets and reported classifier runs.

This distinction permits a precise scientific claim without presupposing the failure mechanism: fixed-skeleton tests leave robustness to changed incidence untested. The tested system might or might not rely on familiar identifiers or motifs. Graph-reasoning studies show sensitivity to graph encoding and structure, which motivates testing these effects separately (Fatemi et al., 2024); they do not verify the practitioner-specific production claim. If future evidence includes operational examples, its access conditions, sampling procedure, and provenance must be documented before claims are expanded.

F.2 Requirements derived for the generation task

The following are our research requirements derived from that motivation, not additional observations attributed to the practitioner.

Controlled variation with preserved evidence. A source, permitted observed context, pinned events, and allowed edit set define a scenario family. The generator should vary unpinned relationships and their continuous attributes without altering the fixed evidence. Distinct incidence patterns and different placements within the observed neighborhood are separate forms of variation. Global identifier renaming and changes in serialization should be controlled rather than counted as new re-

lational scenarios. The implemented menu is bounded; no claim is made that it spans every plausible case or an unseen motif family.

Explicit validity and independently established labels. For a lossless pass-through skeleton, the quantitative criterion is zero net flow at each skeleton-local relay in fixed common units. Ordinary transactions outside the skeleton do not enter the equation, and relay-free cases remain not applicable. This is not a general financial-ledger or power-flow law. Domain rules, a trusted simulator, or independently adjudicated annotations must establish whether a scenario is benign or risky and which entities or events are relevant. An admissible topology and successful DBA-KL completion do not, by themselves, preserve intent, a risk label, or a causal mechanism.

Evaluation access and responsible use. Retrospective testing may permit observations later than a target event while still withholding evaluation labels from parameter fitting. This observation policy must be matched across generators. Benchmark case construction can remain label-informed when explicitly disclosed; it must not be described as a label-blind deployment pipeline. Any operational evaluation requires authorized, isolated testing, appropriate data-access approval, and review of disclosure and release risks. Source-conditioned variants are not automatically anonymized or privacy-preserving. These are conditions for a future study, not assertions that a production intervention has occurred.

Future controlled study. Whether learned structural generation supplies additional valid relational scenarios and independently verified evaluator failures at matched budgets, beyond fixed-topology and simple structural controls, is future work. Its controls coincide with the RQ2 ablations and the RQ3 coverage measurements; in addition it requires positive cases and benign near-matches with independent labels, a tested system that is never its own label adjudicator, fixed prompts and model versions, held-out case parents, and paired, parent-aware uncertainty. A larger number of discovered failures measures test-suite effectiveness, not an improvement of the tested system.

G Related generation and constraint mechanisms

This appendix collects the method-level positioning summarized in Section 1; the concrete adapters used in this manuscript are defined in Appendix C.9.

Joint and temporal graph generation. Existing methods already generate structure and attributes jointly. GDSS jointly diffuses node features and adjacency (Jo et al., 2022); GruM generates structure and features through diffusion mixtures (Jo et al., 2024); and LGD models structure and heterogeneous attributes in a learned latent space (Zhou et al., 2024). DiGress and DeFoG develop discrete diffusion and flow-matching approaches to graph generation, respectively (Vignac et al., 2023; Qin et al., 2025). For temporal interactions, TagGen models temporal-walk structure (Zhou et al., 2020), whereas DG-Gen directly generates event endpoints, times, and features (Hosseini et al., 2025). The present setting additionally fixes selected observations and evaluates topology-dependent positive-flow completions. Joint generation here refers to selection of a topology–attribute pair; the topology is fixed during an individual denoising trajectory.

Constraint satisfaction and completion geometry. ConStruct enforces edge-deletion-invariant structural properties (Madeira et al., 2024); ProbConserv incorporates conservation through a probabilistic update (Hansen et al., 2023); and projected diffusion incorporates constrained optimization into sampling (Christopher et al., 2024). ECI alternates extrapolation, a supplied correction, and interpolation (Cheng et al., 2025). PCFM includes intermediate corrections and a fixed-reference constrained terminal projection (Utkarsh et al., 2025). Suitable adaptations of these methods can satisfy positive relay constraints. Accordingly, the comparison concerns completion objectives and correction schedules, not exclusive access to conservation. The fixed-reference log-distance adaptation and ECI-DBA's changing correction references are specified in Appendix C.9.

Information projection and structural scenarios. DBA-KL builds on information projection and Bregman geometry (Csiszár, 1975; Bregman, 1967). The excess-distortion identity is a property of the KL projection, not an identity exclusive to a particular numerical solver. The contribution uses

signed-incidence analytic updates to compute this reference for boundary-conditioned relay flows and couples completion costs to structural selection. Structural diversity (Velikonivtsev et al., 2024), graph presentation effects (Fatemi et al., 2024), and downstream accuracy remain distinct questions. The practitioner-reported motivation and its evidence boundary are retained in Appendix F; no new production or experimental evidence is asserted here.

Uses of generated interaction data. The introduction cites the application-level motivation: simulator-generated transaction multigraphs for anti-money-laundering benchmarks (Altman et al., 2023), latent-diffusion augmentation for continuous-time dynamic graph models (Tian et al., 2024), graph-diffusion network planning (Kan et al., 2025), power-network operation scenarios (Marot et al., 2021), graph out-of-distribution benchmarks (Gui et al., 2022), counterfactual explanations by structural perturbation (Lucic et al., 2022), motif-aware graph diffusion (Xu & Ma, 2025), discrete-projection diffusion for hard-constrained graph generation (Zhang et al., 2026), and manifold guidance for constrained tabular diffusion (Shankar et al., 2026). These motivate structurally varied, constraint-valid scenarios; none supplies evidence for the practitioner-reported bottleneck or a deployment of the proposed method.

H Extended limitations and scope of the evidence

This appendix expands Section 5. Controls that we did not execute are named as not measured and are never counted as results.

Domain, observation policy, and replication. LI-Small, HI-Small, and LI-Medium are subsets of the same simulator-generated IBM-AML benchmark, so this is one benchmark family rather than several independent application domains, and it contains no production data. The retrospective replacement protocol preserves a shared original background, uses annotation-informed case construction, and permits later-in-window context (Appendix D.2). It tests replacement quality under those conditions; a label-blind, prospective detection procedure is a different setting. Because we fit a single frozen generator per arm and keep stored generation draws, the detector-seed SDs describe detector variability. They say nothing about uncertainty across generator fits, generation draws, or data splits, and the reported aggregates carry no paired covariances or inferential tests (Appendix D.1).

Quantitative validity and structural admissibility. The balance equations cover the modeled events under a lossless, zero-net-accumulation scenario model and a frozen common-unit conversion. They represent a construction rule for generated scenarios, not an account-wide ledger or a general power-flow law. Passing the relay audit therefore speaks to quantitative conservation only; laundering intent, label preservation, and the remaining structural predicates need separate evidence. The flow-corrected terminal Learned populations, for example, contain 29, 90, and 103 closed-observed-account-vocabulary failures even though their quantitative checks pass (Appendix D.6). We keep these failures in the validity and coverage denominators, and quantity completion leaves a vocabulary violation in place.

Restricted structural coverage and selection. The typed grammar and bounded candidate menu explore a restricted neighborhood of the feasible topologies and contain no unseen motif family. Source-relative edits and differences across configurations measure something other than repeated-draw diversity, and the novelty of an emitted structure depends on how we treat identities, pins, directions, multiplicity, and types. A relay-free output carries no conservation equation, so it shows nothing about substantive conservation. The audit finds a higher relay-free share among edited than among fixed-topology outputs (Table 3), and the unweighted ablation of Section 4.3 is less relay-free and emits more, which is consistent with cost weighting favouring relay-free candidates. Because that arm removes correction, feasibility filtering, and weighting jointly, the causal effect of the cost term itself remains open.

Numerical guarantees, accounting, and computational scope. The optimum and the exact identities concern a fixed proposal and a strictly feasible affine problem. Finite tolerances and sweep limits need separate checks, and a failure to converge is a statement about the solver, not about infeasibility. Any accurate solver of the same KL objective shares its optimum and accounting. The

constructed verification in Appendix E addresses the identity and the solvers; an AML detection or production-diagnosis benefit is outside its scope. The cost study times reference implementations on constructed one-to-eight-relay problems. On sufficiently coupled chains DBA-KL is slower than the same-KL Newton solver, and fewer correction calls on their own say little about whole-generation speedup (Appendix D.7). End-to-end denoising, candidate construction, selection, auditing, persistence, and a hybrid Newton fallback stay outside the measurement: the reported costs are per-call solver costs.

Application validity and operational evidence. The industrial motivation is a qualitative practitioner report. It provides neither a production deployment nor a measured LLM-robustness result. The controlled reasoning study in Appendix F.2, independently established scenario labels, and transfer to a smart-grid setting all belong to future work. Source-conditioned variants retain observations and context, so they are not automatically anonymized or privacy-preserving. Operational use would need the independent domain validation and authorized evaluation described in Appendix F.