FROM SEQUENCE UNCERTAINTY TO PHENOTYPE HYPOTHESES: UNCERTAINTY-CONDITIONED INFERENCE OVER BIOMEDICAL KNOWLEDGE GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models provide variant-level evidence, but connecting that evidence to phenotype hypotheses requires accounting for uncertainty throughout structured reasoning. We introduce an uncertainty-conditioned path-energy framework that converts sequence predictions and calibrated predictive uncertainty into explicit distributions over graph-supported phenotype hypotheses, without patient-level phenotype supervision. An interpretable learned energy scores admissible gene–disease–phenotype routes, query uncertainty controls the target temperature, and path probabilities are aggregated into phenotype hypotheses. This formulation makes evidence quality an explicit determinant of hypothesis breadth while retaining traceable multi-path support. Exact normalization over enumerable candidate sets provides a reference for evaluating a Hastings-corrected learned revision operator. On AlzKB and Hetionet, phenotype-level entropy tracks upstream uncertainty, with Spearman correlations of 0.951 and 0.878, respectively; ablations identify the contribution of uncertainty conditioning. In masked-edge recovery, the model exceeds chance and degree-matched probability-permutation controls. With query genes held out from calibration and energy learning, mean recall@10 reaches 0.206 across three edge masks, exceeding DistMult by 0.108, with a roughly fivefold lower across-mask coefficient of variation. These results establish an uncertainty-aware interface from genomic predictions to graph-supported phenotype hypotheses that combines interpretable probability allocation, improved held-out association recovery, and reduced sensitivity to graph masking.

1 INTRODUCTION

Interpreting a genetic variant requires connecting molecular evidence to the phenotypes it could plausibly explain. Genomic sequence models provide variant-level predictions (Nguyen et al., 2024; Brixi et al., 2026), while biomedical knowledge graphs connect genes to diseases and phenotypes through multiple evidence routes. The quality of the upstream evidence should influence how probability is distributed across these competing explanations: confident evidence can support a concentrated set of hypotheses, whereas uncertain evidence calls for a broader allocation. A point prediction alone does not specify this allocation.

Existing work supplies complementary components. Gene–phenotype language models answer genetics questions (Lyu et al., 2024), database-grounded agents generate and verify functional descriptions (Wang et al., 2025), and graph reasoners retrieve interpretable evidence paths (Luo et al., 2024; Sun et al., 2024). We address the interface between upstream evidence and downstream hypotheses: given a sequence prediction and its predictive uncertainty, how should the model allocate probability across graph-supported phenotype explanations?

We formulate this problem as uncertainty-conditioned inference over admissible paths. Given a query gene, a sequence prediction p, and calibrated uncertainty H, an interpretable learned utility U_w defines the target $\pi_H(\pi) \propto \exp\{U_w(\pi)/\theta(H)\}$. Aggregating path mass yields a normalized distribution over phenotype hypotheses. The learned energy determines the relative support for competing routes,

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.

while the uncertainty-dependent temperature controls their concentration. Enumerable candidate supports allow exact evaluation of the target and direct verification of a Hastings-corrected learned revision operator (Section 3.4).

We make three contributions. **Uncertainty-conditioned phenotype inference.** We introduce a query-conditioned probabilistic interface that carries genomic predictive uncertainty into a structured distribution over phenotype hypotheses, connecting evidence quality to hypothesis breadth. **Interpretable and auditable structured inference.** We learn interpretable scores over graph routes, aggregate evidence for each phenotype, and verify the sampling procedure against the exact target distribution. **Improved recovery and robustness.** Across AlzKB and Hetionet, we separately evaluate uncertainty response, held-out association recovery, and sampling accuracy. In gene-held-out evaluation across three graph masks, our method improves mean recall@10 over DistMult by 0.108 while reducing across-mask variability from 0.49 to 0.10. Matched controls further show that recovery depends on where the model places probability mass.

2 Related work

Genomic prediction and phenotype interpretation. Genomic sequence models provide variant-effect predictions (Nguyen et al., 2024; Brixi et al., 2026). Gene–phenotype language models such as GP-GPT answer genetics questions and determine relations in text (Lyu et al., 2024), while database-grounded agents generate and verify functional descriptions of gene sets (Wang et al., 2025). Our framework connects these levels of interpretation through an explicit probabilistic interface: a variant-level prediction and its calibrated uncertainty determine a distribution over graph-supported phenotype hypotheses. The central object is the allocation of probability across competing explanations, with upstream evidence quality controlling the breadth of that allocation.

Graph-supported reasoning and explanation. Path-based reasoners learn graph traversal policies (Xiong et al., 2017; Das et al., 2017), RoG retrieves evidence through planned relation paths (Luo et al., 2024), and ToG searches a graph to support an answer (Sun et al., 2024). Embedding methods and relational graph networks score graph relations (Bordes et al., 2013; Yang et al., 2014; Schlichtkrull et al., 2018), while PageRank-style propagation and relation-path features provide complementary graph-based mechanisms (Page et al., 1999; Lao et al., 2011). Our formulation makes the query-conditioned distribution over graph-supported hypotheses an explicit inference output, evaluated both for uncertainty response and for recovery of masked associations. Exact normalization over admissible paths and aggregation of multi-route evidence make this output directly inspectable. Comparisons with DistMult and PageRank use identical candidate supports, and the gene-held-out, mask-averaged evaluation demonstrates improved recovery over DistMult (Section 5.7).

Probabilistic and uncertainty-aware reasoning. DeepProbLog integrates neural probabilities into probabilistic logic programs (Manhaeve et al., 2018), UKGE models confidence in graph facts (Chen et al., 2019), and BetaE represents logical queries with distributions whose entropy reflects answer-set structure (Ren & Leskovec, 2020). Agent-level methods propagate uncertainty along reasoning trajectories (Zhao et al., 2025), while predictive uncertainty estimation, calibration, and selective prediction characterize a predictor's reliability (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Guo et al., 2017; Geifman & El-Yaniv, 2017). We use predictive uncertainty supplied by an upstream sequence model to condition a downstream phenotype distribution. This couples uncertainty about the input evidence to the concentration of alternative graph-supported explanations, complementing approaches that model confidence in graph facts or assess the reliability of a reasoning output.

Structured distributions and sampling. Finite energy models, temperature-controlled Gibbs distributions, and Metropolis–Hastings transitions provide the probabilistic foundation for our construction (Metropolis et al., 1953; Hastings, 1970). GFlowNets learn to sample structured objects in proportion to reward, including temperature-conditioned variants with controllable diversity (Bengio et al., 2021; Kim et al., 2024). We use temperature as an interface to query-specific predictive uncertainty, making hypothesis breadth depend on evidence quality rather than solely on a diversity setting. Enumerable path supports additionally permit exact normalization and direct measurement of sampling error.

This separates the behavior and predictive utility of the learned distribution from the fidelity of the operator used to sample it.

3 METHOD

3.1 PROBLEM SETUP AND VALID SUPPORT

Let $G = (V, E)$ be a directed typed graph and g a start node. The finite support $\mathscr{C}(g)$ contains complete valid paths ending in a designated outcome type. Every consecutive node pair must be an observed graph edge or a documented semantic inverse. For each variant, the upstream sequence model supplies a prediction $p \in [0,1]$ calibrated using definitive Pathogenic/Benign labels, together with a calibrated uncertainty score $H \in [0,1]$. Here, p summarizes the model's evidence about variant pathogenicity, while higher H indicates greater uncertainty in that evidence.

For AlzKB, admissible phenotype-reaching paths follow the template

$$\text{Gene}_0 \to [\text{Gene}_1] \to \text{Disease} \to \text{Phenotype},$$

where the gene relay is optional and uses a typed interaction, regulation, or co-variation edge. We retain all direct routes and at most 24 deterministic, relation-stratified route prefixes. BiologicalProcess and Pathway neighbors annotate route genes but cannot be traversed to phenotypes in this release.

3.2 INTERPRETABLE LEARNED ENERGY

Each path has a 14-dimensional feature vector containing relation-frequency structure, length, gene-edge type, disease coverage, endpoint specificity, endpoint-group indicators and crowding, route-gene mechanism annotations, and interactions with p and $1-H$. Standardized features define $U_w(\pi) = w^\top \widetilde{\phi}(\pi)$. For labeled query v, let c_i mark a prespecified core outcome and $m_v = \sum_i c_i \pi_H(\pi_i)$. A logistic readout predicts $\hat{y}_v = \sigma(\alpha m_v + \beta)$. We optimize

$$\frac{1}{N}\sum_v \text{BCE}(y_v, \hat{y}_v) + \frac{\lambda}{2}\|w\|_2^2 + \frac{\lambda_\alpha}{2}\alpha^2,$$

using bounded L-BFGS, $\lambda = 10^{-2}$, and $\lambda_\alpha = 10^{-3}$. Only 121 definitive Pathogenic or Benign variants (Richards et al., 2015) enter sequence calibration or energy learning.

3.3 UNCERTAINTY-CONDITIONED TARGET

One schedule defines the target:

$$\theta(H) = 0.35 + 0.90H, \qquad \pi_H(i) = \frac{\exp\{U_i/\theta(H)\}}{\sum_j \exp\{U_j/\theta(H)\}}.$$

We obtain each phenotype's probability by summing the probabilities of all paths ending at that phenotype, and group probabilities by summing across phenotypes within each group, without applying another softmax. These probabilities express relative model support among graph-supported hypotheses for a given variant, not the clinical probability that a patient has or will develop a phenotype.

Proposition 1 (Fixed-energy entropy monotonicity)**.** *For fixed finite utilities U, the Shannon entropy of* $\text{softmax}(U/\theta)$ *satisfies* $\frac{d\mathscr{H}}{d\theta} = \text{Var}_{\pi_\theta}(U)/\theta^3 \geq 0$. *Thus any non-decreasing $\theta(H)$ produces non-decreasing path entropy at fixed U.*

This result characterizes path-level concentration at fixed utilities. We evaluate phenotype-level breadth after path aggregation, including the effects of uncertainty-dependent utility features, in the experiments below.

3.4 HASTINGS-CORRECTED LEARNED REVISION

We use the full-support independence proposal $q_H(i) = 0.90\,\text{softmax}_i\{U/(2\theta(H))\} + 0.10/|\mathscr{C}(g)|$. From current state i, propose $j \sim q_H$ and accept with

$$A(i,j) = \min\left\{1, \exp\left(\frac{U_j - U_i}{\theta(H)}\right)\frac{q_H(i)}{q_H(j)}\right\}.$$

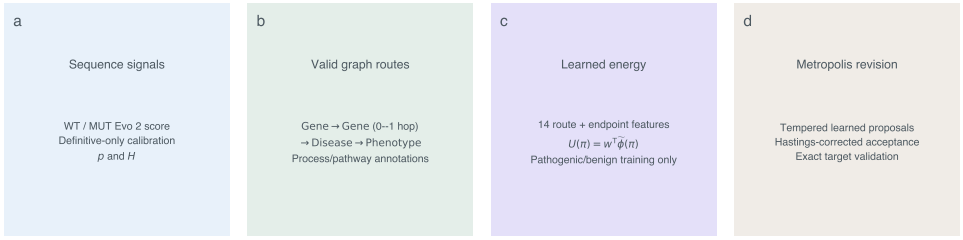


Figure 1: Uncertainty-conditioned path-energy inference. Sequence predictions and calibrated uncertainty condition an interpretable energy over valid zero- or one-gene-hop routes. Path probabilities are aggregated into phenotype hypotheses, and a full-support learned/uniform revision proposal is accepted with the Hastings correction and evaluated against the exact target.

Proposition 2 (Target invariance)**.** *The transition kernel above is irreducible, aperiodic, and reversible with respect to π_H; consequently π_H is stationary.*

Full support follows from the uniform component, and detailed balance follows from the standard MH ratio. We initialize at the first deterministically sorted shortest path and run $50 + 10\lceil \log_2 |\mathscr{C}(g)| \rceil + \text{round}(25H)$ revisions in each of 512 independent chains. The measured mixing curve (§5.1) shows the chain reaches its Monte-Carlo floor within about ten revisions; the longer schedule is a conservative margin, not a requirement. Main distributional statistics use the exact finite target; chain outputs are used to test implementation fidelity and final-state coverage.

Exact-reference verification of learned revision. Enumerable candidate supports make the target directly computable, allowing the revision operator to be evaluated against an exact reference. We use the exact target for distributional statistics and the chain outputs to assess sampling fidelity and coverage. The Hastings factor preserves the specified uncertainty-conditioned distribution under the learned proposal (Metropolis et al., 1953; Hastings, 1970). In particular, an independence proposal q_H combined with an uncorrected Metropolis acceptance ratio generally has stationary mass proportional to $\pi_H(i)\, q_H(i)$ rather than $\pi_H(i)$; the correction removes this proposal-induced distortion, which would otherwise sharpen the distribution and undo part of the uncertainty conditioning. The resulting operator provides a verified sampling implementation for the formulated target. On the finite supports studied here, direct categorical sampling is also available; the purpose of the chain experiments is to establish fidelity, not a runtime improvement over exact sampling.

4 Experimental design

Case study. We use 985 Evo 2-scored ClinVar (Landrum et al., 2018) variants matched to AlzKB (Romano et al., 2024); 982 have a valid path. 3 variants from GRID2IP, PLCD1, SORCS1 are excluded for absent routes. Each query corresponds to one variant. Variants in the same gene share the candidate graph paths, but their individual p and H values can produce different probability distributions over phenotypes. The 982 variants span 15 genes, with APP, PSEN1, and PSEN2 accounting for 94% of queries; confidence intervals therefore resample genes as clusters. Candidate sets contain at most 20 route prefixes and 1060 complete paths. Sequence probabilities and isotonic uncertainty (Zadrozny & Elkan, 2002) are calibrated using definitive labels only.

Comparisons on identical supports. Component controls use the same support and learned weights: uniform path mass; fixed temperature $\theta = 0.80$; removal of confidence interactions; both removals; and removal of route mechanism features. Non-learned baselines are personalized PageRank from the start gene on the retained graph ($\alpha = 0.85$, restricted to reachable endpoints and renormalized) and beam truncation, which places uniform mass on the top-k paths under the learned utility without an uncertainty-scaled temperature ($k \in \{1, 5, 20\}$). These are component controls, not competitive baselines; the trained comparator is the DistMult reference of Section 5.7.

Sampler checks. Mixing is measured by running 2,048 independent chains per query and recording the empirical-to-exact divergence at fixed step counts (1–100), at the paper's schedule, at twice

the schedule, and at the earlier $10 + \text{round}(15H)$ schedule; chain-count sensitivity uses the first $N \in \{32, 128, 512, 2048\}$ chains. The synthetic check generates heterogeneous utilities over ten groups in spaces of 53, 256, and 1,024 states with five uncertainty levels, 30 replicates, and 512 chains.

Second graph. Hetionet v1.0 (Himmelstein et al., 2017) is expressed in the same schema (34,723 nodes, 1,341,335 edges over gene, disease, symptom, process and pathway types) so that enumeration, features, learning, target and sampler run unchanged. Its 3,357 disease–symptom edges leave 133 diseases, and 373 of the 415 linked symptoms have several disease parents (up to 49), so endpoint availability is gene-dependent. The same 985 variants keep their frozen (p, H) and labels; all 18 genes have valid routes (13–967 candidate paths, 13–330 endpoints). A knowledge graph without patient outcomes offers one source of ground truth: its own edges. We therefore test predictive generalization by masking 20% of disease–symptom edges (671 of 3,357), rebuilding supports and retraining the energy on the reduced graph, and asking whether the path distribution places mass on true outcomes it never saw—held-out symptoms of each gene's own diseases that remain reachable on the reduced graph (recall@k, MRR; 964 variants in 15 genes). The ranking borrows the filtered protocol of knowledge-graph completion, but the object evaluated is the query-conditioned path distribution, not a triple scorer. Filtering removes every symptom still attached to the gene's own diseases, so a recoverable held-out symptom is reached only through gene relays $g \to g' \to d' \to s$ (worked query: Appendix H). A trained DistMult embedding (Yang et al., 2014) (dimension 32, logistic loss, Adagrad, two-hop composition) is the reference for a trained triple scorer on the same candidates; it receives no (p, H). Two prespecified tests follow, with recall@10 primary. *Matched controls* keep each query's candidates, positives and full-model probabilities and compare against chance, symptom popularity (degree in the reduced graph), and a degree-matched permutation that shuffles the model's probabilities among candidates of similar degree (one permutation per gene, 200 draws), preserving values and entropy but not their assignment. Variants in the same gene share the same reachable held-out symptoms; recovery metrics average over variant queries unless labeled gene-balanced (equal weight per gene). Gene-held-out evaluation excludes all variants in one group (APP, PSEN1, PSEN2, or the remaining genes) from calibration and energy fitting, using graph-mask seeds 42, 7, and 123. The held-out genes remain in the graph used for inference. A noninferiority margin of 0.02 against DistMult was fixed in advance; leakage and coverage checks are in App. F.

Energy-recovery study. Queries are generated with K candidate paths, 14-dimensional standardized features including a binary core-outcome indicator, a planted energy w^*, and labels drawn from the same readout family as the model. The exact training routine of the paper is applied for $K \in \{53, 256, 1024\}$ and $N \in \{121, 500, 2000\}$ over five replicates. Recovery is measured on 500 held-out queries by weight cosine, utility rank agreement, group-level JS between learned and planted targets, and label AUROC against the oracle readout.

5 Results

5.1 The implemented chain samples the stated target

Across 982 real queries, the sampled phenotype-group distribution closely matches the exact target, with mean Jensen–Shannon divergence (JS) (Lin, 1991) of 0.0031 (median 0.0029) and mean acceptance rate of 0.507 (Fig. 2a). Increasing the number of chains from 32 to 128, 512, and 2,048 reduces JS from 0.043 to 0.012, 0.0033, and 0.0008, respectively. This approximately $1/N$ decrease indicates that the remaining difference is consistent with finite-sample error rather than persistent sampling bias. Convergence is also rapid: JS falls from 0.155 after one revision to 0.053, 0.0070, and 0.0034 after two, five, and ten revisions, then levels off (Fig. 2b); both the main schedule (mean 153 steps) and the shorter one (mean 16 steps) sit on this floor. Synthetic experiments show similar agreement across uncertainty levels and state spaces of up to 1,024 states: the largest condition-level mean JS is 0.0038 and the largest 95th-percentile JS is 0.0069, with little variation across H or state-space size (Fig. 6a). Together, these results verify the sampling implementation and distinguish finite-sample error from the behavior of the target distribution, which we evaluate next.

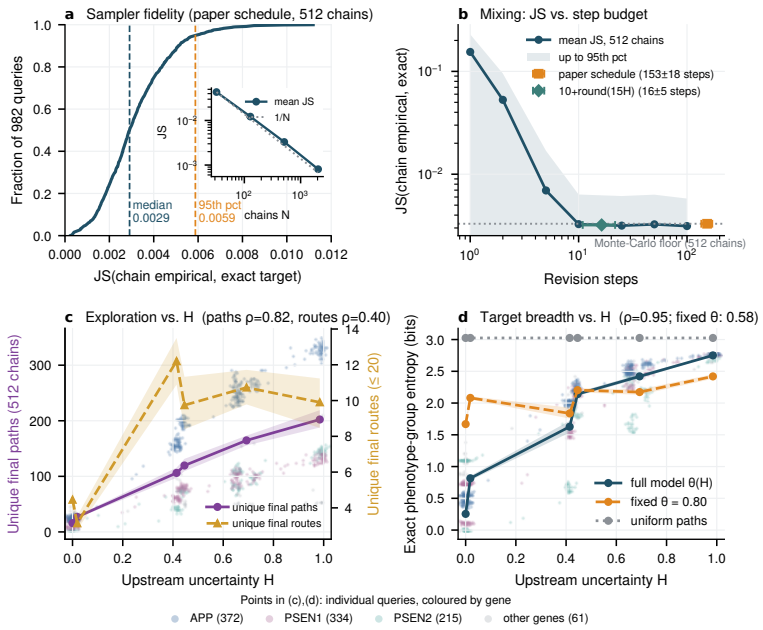


Figure 2: Sampler fidelity and uncertainty behavior on 982 queries (AlzKB). (a) Empirical cumulative distribution of chain-to-target JS at the paper's schedule; inset: mean JS versus number of chains with a $1/N$ reference. (b) Mixing curve: mean JS (shaded to the 95th percentile) versus revision steps; the paper's and the earlier schedule are marked (legend top right). (c) Unique final paths (left axis, purple) and routes (right axis, gold, capped at 20) across 512 chains versus H; gene-cluster 95% CIs 0.79–0.96 (paths), 0.10–0.78 (routes). (d) Exact phenotype-group entropy versus H for the full model, fixed θ, and uniform paths (legend bottom right); cluster CIs 0.90–0.98 (full), 0.47–0.85 (fixed θ). Curves are quantile-binned means with 95% bootstrap bands. Small points in (c),(d) are individual queries coloured by gene (key with counts below the panels), jittered in H because the isotonic calibrator is a step function; the vertical spread within an H plateau is gene-dependent support.

5.2 Upstream uncertainty controls structured hypothesis breadth

Phenotype hypotheses broaden as uncertainty increases. From the lowest to the highest uncertainty quantile, phenotype-group entropy rises from 0.585 to 2.631 bits, while the probability assigned to the top-ranked group falls from 0.864 to 0.357. Across variants, entropy correlates strongly with H at both the group level ($\rho = 0.951$, gene-cluster 95% CI 0.898–0.984) and the individual-phenotype level ($\rho = 0.890$, 0.852–0.926). Final states from 512 chains also become more diverse with higher H, with correlations of $\rho = 0.821$ (0.792–0.961) for paths and $\rho = 0.403$ (0.097–0.775) for routes. The cap of 20 routes per gene limits route-level variation (Fig. 2c,d). Recomputing the target with kNN-calibrated uncertainty instead of isotonic calibration preserves a strong entropy–uncertainty correlation ($\rho = 0.869$, 0.777–0.964), showing that the relationship holds under both calibrators. The following ablations examine how explicit uncertainty conditioning contributes to this relationship, while the masked-edge experiments test whether the resulting probability assignments help recover held-out associations.

5.3 Ablations separate breadth control from probability allocation

Table 1 and Fig. 3 separate the effects of uncertainty conditioning and learned route features. Fixing the temperature reduces the correlation between upstream uncertainty and phenotype-group entropy from 0.951 to 0.582, and removing both the temperature schedule and confidence interactions gives 0.605 (intervals in Table 1): these controls attenuate the strong uncertainty–breadth relationship of the full model. Removing mechanism features changes the target distribution (mean JS 0.073) while retaining a high entropy correlation, showing that route annotations affect probability allocation beyond the temperature schedule. Uniform and personalized PageRank propagation exhibit near-zero

Table 1: Component controls and baselines on identical candidate supports (982 queries). Spearman ρ is between H and grouped entropy; brackets are gene-cluster 95% bootstrap intervals over 15 genes.

Model	Entropy	Top-1	JS to full	ρ [95% CI]
Uniform paths	3.025	0.264	0.287	−0.005 [−0.03, 0.01]
Personalized PageRank	3.025	0.264	0.287	−0.030 [−0.15, 0.03]
Beam $k = 20$ (no $\theta(H)$)	2.060	0.414	0.186	−0.273 [−0.58, 0.28]
Beam $k = 5$ (no $\theta(H)$)	0.840	0.726	0.158	−0.027 [−0.20, 0.23]
Fixed θ + no confidence interactions	2.040	0.508	0.054	0.605 [0.50, 0.90]
Fixed temperature $\theta = 0.80$	2.078	0.515	0.052	0.582 [0.47, 0.85]
No mechanism features	2.127	0.458	0.073	0.913 [0.90, 0.93]
No confidence interactions	1.695	0.589	0.002	0.944 [0.89, 0.99]
Full model (reference)	1.662	0.606	0.000	0.951 [0.90, 0.98]

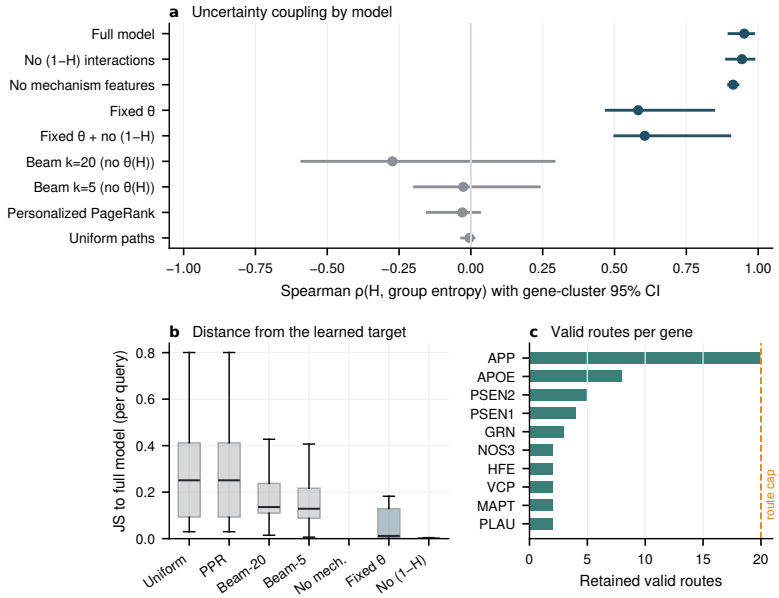


Figure 3: Controls and baselines (AlzKB). (a) H–entropy association with gene-cluster 95% CIs; navy points are component controls, grey points are non-learned or truncation baselines (no separate legend; the row labels name each model). (b) Per-query JS to the full model, one box per model. (c) Retained valid routes per gene; the dashed line marks the route cap of 20. Every route reaches the same 53 endpoints through one disease node, so uniform and PageRank propagation coincide.

associations with uncertainty, and beam truncation under the learned utility does not reproduce the full model's relationship. Together, the controls distinguish the uncertainty-dependent breadth of the target from the route features that shape its allocation.

5.4 Usable graph topology shapes hypothesis support

AlzKB contains 210,656 retained nodes and 1,199,652 edges, but its phenotype-reaching subgraph has a specific structure: all 53 `DISEASEMANIFESTSSYMPTOM` edges originate from one disease node, and Pathway and BiologicalProcess nodes provide annotations rather than additional phenotype-reaching continuations. Consequently, every evaluated gene has the same 53 phenotype endpoints, although the routes supporting those endpoints differ. Under the evaluated propagation baselines, personalized PageRank and uniform path mass coincide (Table 1). The learned energy can nevertheless assign different weights to the valid evidence routes. Hetionet (Section 5.6) provides a complementary evaluation in which multiple disease parents and gene-dependent endpoint availability test the same

framework on a different reasoning topology. Perturbation sensitivity of the target is reported in App. E.

5.5 Energy recovery and optimization diagnostics

The synthetic study tests whether the model can recover a known path utility used to generate the data. With at most 256 candidate paths and at least 500 training queries, successful fits achieve weight cosine similarity of 0.98–0.99 and held-out AUROC within 0.01 of the oracle (Fig. 6d). On both real graphs, repeated fits produce identical objective values and weight cosine similarity of 1.000 across the tested restarts and tolerance settings. Some synthetic fits collapse when labeled data are limited or candidate sets are large. App. D provides the full recovery results, failure rates, and analysis of initialization choices.

5.6 Uncertainty-conditioned breadth replicates across graph topologies

On Hetionet, the energy trains reproducibly using the same 121 definitive anchors, reaching objective 0.238 compared with 0.640 for a constant predictor (App. D). Endpoint entropy of the exact target increases from 2.020 to 6.413 bits across uncertainty strata while top-1 mass decreases from 0.529 to 0.052. The uncertainty–entropy correlation is 0.878 (gene-cluster 95% CI 0.780–0.910), compared with 0.495 (0.281–0.797) under fixed temperature (Table 3, Fig. 5b). Unlike AlzKB, Hetionet yields distinct uniform and PageRank distributions, but neither exhibits a meaningful association with uncertainty ($\rho = 0.027$, 0.028). Final-state diversity also tracks uncertainty ($\rho = 0.896$), and the sampler achieves mean grouped JS of approximately 0.003 on supports averaging 713 states (Fig. 5c; further statistics in App. C). These results reproduce uncertainty-conditioned hypothesis breadth under gene-dependent phenotype availability.

5.7 Improved recovery and lower across-mask variability

Gene-held-out recovery. With each query gene excluded from sequence calibration and energy learning, the framework achieves mean recall@10 of 0.206 across three edge masks (Table 2, Fig. 4c). The paired improvement over DistMult is 0.108 (gene-cluster 95% CI 0.024–0.160), establishing higher aggregate recovery in this evaluation. The observed mean improvement is positive on all three masks: 0.032 for seed 42 (95% CI -0.008 to 0.041), 0.156 for seed 7 (0.012–0.264), and 0.137 for seed 123 (0.037–0.227).

Reduced sensitivity to graph masking. Recall@10 spans 0.182–0.222 across masks for the proposed method against 0.057–0.151 for DistMult, a coefficient of variation of 0.10 versus 0.49 (roughly fivefold lower relative variability); the method thus combines higher mean recovery with more consistent performance across the three evaluated masks.

Recovery beyond matched controls. On the primary masked graph, recall@10 is 0.163, exceeding chance at 0.063 (paired difference 0.100, 95% CI 0.034–0.122) and a degree-matched permutation of the model's probabilities at 0.072 (paired difference 0.092, 95% CI 0.011–0.136; $p < 0.005$; Fig. 4b,d). The permutation preserves the candidate set, positives, probability values, and entropy while changing their assignment. This comparison shows that recovery depends on where the model places probability mass, rather than on distributional breadth alone.

Primary split and conditioning ablations. On the primary split, the full model achieves recall@10/recall@25/MRR of 0.163/0.364/0.067, compared with 0.151/0.329/0.061 for DistMult (Table 3, Fig. 4a). These point estimates favor the proposed method; the paired intervals on that individual split include zero ($+0.013$, CI -0.044 to 0.054). Fixed-temperature ablations attain similar recall, with full-minus-fixed recall@10 differences of 0.011 (95% CI -0.011 to 0.033) on the primary split and -0.015 (-0.026 to 0.001) under gene hold-out. The recovery comparison thus evaluates the full structured-inference framework, while the breadth experiments identify the contribution of uncertainty conditioning. Recovery is evaluated over reachable masked symptoms on matched candidate supports; coverage, per-gene and popularity comparisons, and the prespecified single-split noninferiority analysis are in App. F, and the descriptive recall–H association in App. G.

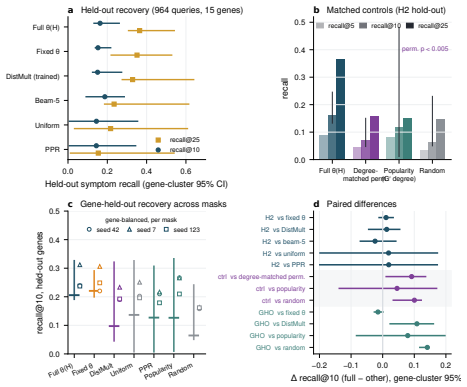


Figure 4: Masked-edge recovery (Hetionet). (a) Recall of masked symptoms on the primary hold-out, recall@10 (navy) and recall@25 (gold), gene-cluster 95% CIs. (b) Matched controls on identical candidates and positives; shade encodes recall@5/10/25, bars are recall@10 CIs. (c) Gene-held-out recovery: pooled mean with gene-cluster 95% CI over three masks; open markers are gene-balanced per-mask means. (d) Paired recall@10 differences, full minus each alternative: primary hold-out (navy), matched controls (purple), gene hold-out (teal).

Table 2: Gene-held-out recovery across three edge masks (Hetionet; query genes withheld from calibration and energy learning). Pooled means are query-weighted over 3×964 queries with gene-cluster 95% CIs; range and coefficient of variation (CV) are over the three per-mask means.

recall@10	Full, $\theta(H)$	DistMult
Pooled mean [CI]	0.206 [0.19, 0.33]	0.097 [0.05, 0.32]
Per-mask range	0.182–0.222	0.057–0.151
Across-mask CV	0.10	0.49
Full − DistMult [CI]	+0.108 [0.024, 0.160]	

Table 3: Hetionet (985 queries, 18 genes; masked-edge recovery on 964 queries, 15 genes). ρ is between H and endpoint entropy (gene-cluster 95% CI) and supports the breadth claim; recall and MRR are conditional on reachable masked symptoms (primary split, Section 5.7).

Model	ρ [CI]	r@10	r@25	MRR
Full model, $\theta(H)$	0.878 [0.78, 0.91]	0.163	0.364	0.067
Fixed temperature	0.495 [0.28, 0.80]	0.153	0.351	0.065
DistMult (trained)	–	0.151	0.329	0.061
Beam $k = 5$	0.014 [−0.07, 0.10]	0.187	0.235	0.050
Uniform paths	0.027 [−0.11, 0.09]	0.145	0.217	0.077
PageRank (personalized)	0.028 [−0.12, 0.09]	0.144	0.154	0.054

6 Secondary diagnostic: variant classification

Variant classification evaluates a different output from phenotype-hypothesis recovery. Across five outer folds, graph-augmented AUROC is 0.968 compared with 0.955 for sequence-only features, with a paired difference of 0.0126 (gene-cluster 95% CI −0.0235 to 0.0191). Brier score and expected calibration error are higher with graph features. We report this diagnostic separately from the recovery results; full metrics and the locked-label protocol appear in App. I.

7 Future work

Future work will extend the evaluation to more genes and biomedical graphs, compare with relational and path-policy reasoners, and improve learning with larger candidate sets and limited labeled data. Patient-level studies could establish how the output distributions relate to observed phenotypes and investigate their calibration for clinical risk estimation.

8 Conclusion

We introduced an uncertainty-conditioned path-energy framework that connects genomic sequence predictions to explicit distributions over graph-supported phenotype hypotheses. The framework combines interpretable route utilities, query-dependent control of hypothesis breadth, and exact-reference validation of learned revision. Across two biomedical knowledge graphs, hypothesis breadth tracks upstream uncertainty, and masked-edge tests show that the learned probability allocation recovers held-out associations beyond chance and matched permutation controls. With query genes held out from calibration and energy learning, the method improves mean recall@10 over DistMult by 0.108 and achieves a roughly fivefold lower across-mask coefficient of variation. Together, these results advance uncertainty-aware genotype-to-phenotype interpretation through interpretable inference, improved association recovery, and greater consistency across the evaluated graph masks.

REPRODUCIBILITY STATEMENT

All reported numbers are generated from frozen experiment tables and checked programmatically against the manuscript. Split manifests, model hashes, and per-experiment logs document the calibration, training, and evaluation pipeline. Reproducibility checks cover path validity, feature dimensions, detailed balance, empirical stationarity, full proposal support, empty and singleton supports, deterministic seeds, and mass conservation. Anonymized code, frozen result tables, figure scripts, and the manuscript-number verifier are available online.[1]

AI USE STATEMENT

AI assistants were used to polish the writing, to audit the writing for consistency between the reported numbers, tables, and claims, and to audit the code. The authors reviewed all AI-assisted content, verified every number and claim against the frozen experiment outputs, and take full responsibility for the final manuscript and code.

REFERENCES

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3363–3370, 2019.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6:e26726, 2017.

Minsu Kim, Joohwan Ko, Taeyoung Yun, Dinghuai Zhang, Ling Pan, Woo Chang Kim, Jinkyoo Park, Emmanuel Bengio, and Yoshua Bengio. Learning to scale logits for temperature-conditional GFlowNets. In *International Conference on Machine Learning (ICML)*, pp. 24248–24270, 2024.

[1] `https://anonymous.4open.science/r/Uncertainty-Conditioned-Inference-00E1`

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

Ni Lao, Tom Mitchell, and William W Cohen. Random walk inference and learning in a large scale knowledge base. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 529–539, 2011.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14400–14423, 2024.

Yanjun Lyu, Zihao Wu, Lu Zhang, Jing Zhang, Yiwei Li, Wei Ruan, Zhengliang Liu, Zeyu Zhang, Xiang Li, Rongjie Liu, et al. Gp-gpt: Large language model for gene-phenotype mapping. *arXiv preprint arXiv:2409.09825*, 2024.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*, 31, 2018.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. *Advances in Neural Information Processing Systems*, 33:19716–19726, 2020.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, 17(5):405–423, 2015.

Joseph D Romano, Van Truong, Rachit Kumar, Mythreye Venkatesan, Britney E Graham, Yun Hao, Nick Matsumoto, Xi Li, Zhiping Wang, Marylyn D Ritchie, et al. The alzheimer's knowledge base: a knowledge graph for alzheimer disease research. *Journal of Medical Internet Research*, 26: e46777, 2024.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, volume 2024, pp. 3868–3898, 2024.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025.

Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 564–573, 2017.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. Uncertainty propagation on LLM agent. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6064–6073, 2025.

A PROOFS

Entropy monotonicity. Let $Z(\theta) = \sum_i e^{U_i/\theta}$ and $\pi_i = e^{U_i/\theta}/Z$. Since $\mathscr{H} = \log Z - \mathbb{E}_\pi[U]/\theta$, differentiation yields $d\mathscr{H}/d\theta = \mathrm{Var}_\pi(U)/\theta^3 \geq 0$.

MH invariance. For $i \neq j$, $\pi(i)q(j)A(i,j) = \min\{\pi(i)q(j), \pi(j)q(i)\} = \pi(j)q(i)A(j,i)$. The uniform mixture makes $q(j) > 0$ for every state; rejection provides self-loops. The finite chain is therefore irreducible, aperiodic, reversible, and stationary for π.

B INTERVAL ESTIMATES

Table 4 lists every association reported in the main text with naive variant-level and gene-cluster bootstrap intervals (1,000 replicates). The cluster intervals are wider because three genes carry 94% of variants.

Table 4: Spearman associations with H (isotonic unless stated).

Quantity	ρ	IID 95% CI	Gene-cluster 95% CI
Grouped phenotype entropy	0.951	[0.943, 0.959]	[0.898, 0.984]
Raw endpoint entropy	0.890	[0.878, 0.902]	[0.852, 0.926]
Top-1 group mass	−0.904	[−0.918, −0.890]	[−0.960, −0.701]
Unique final paths (512 chains)	0.821	[0.793, 0.842]	[0.792, 0.961]
Unique final routes (512 chains)	0.403	[0.345, 0.458]	[0.097, 0.775]
Acceptance rate	0.492	[0.428, 0.564]	[−0.030, 0.801]
Perturbation JS (±0.30)	−0.677	[−0.718, −0.638]	[−0.818, −0.276]
Perturbation top-1 switch	0.351	[0.306, 0.394]	[0.006, 0.467]
Grouped entropy, kNN-recomputed target	0.869	[0.847, 0.888]	[0.777, 0.964]
H_{iso} vs. $H_{k\mathrm{NN}}$	0.657	[0.612, 0.697]	[0.636, 0.783]

C SECOND-GRAPH PANELS

Additional Hetionet statistics referenced in Section 5.6: anchor AUROC of the trained energy 0.973; exact-target endpoint entropy 7.655 bits under uniform path mass and 7.252 under personalized PageRank; final-state path diversity across 512 chains $\rho = 0.896$ (gene-cluster 95% CI 0.794–0.944); and the single-split label-prediction check on this graph, a paired AUROC difference of −0.013 (CI −0.136–0.011), which is not part of the recovery evidence (Section 6).

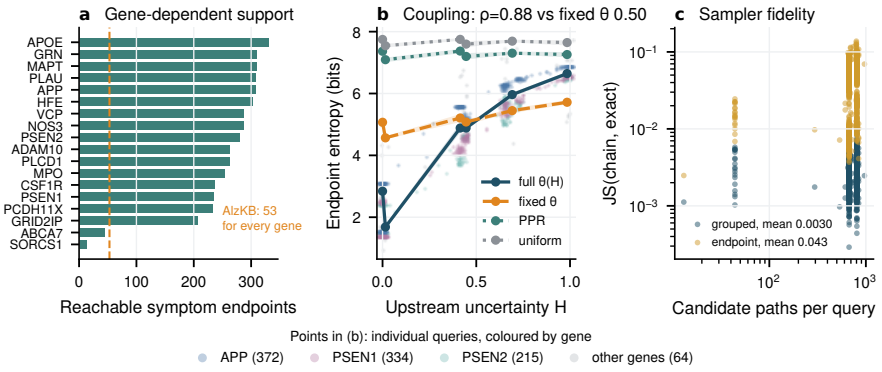


Figure 5: Hetionet. (a) Reachable endpoints per gene (AlzKB gives 53 for every gene, dashed line). (b) Endpoint entropy versus H for the full model, fixed θ, PageRank and uniform (line legend bottom right); small points are full-model queries coloured by gene, with the colour key and per-gene counts below the panels. (c) Chain-to-target JS versus support size on log axes, grouped and endpoint-level (legend bottom left with means).

D ENERGY-RECOVERY AND OPTIMIZATION DIAGNOSTICS

With the paper's training routine, successful fits recover the planted energy closely: for $K \leq 256$ and $N \geq 500$, weight cosine is 0.98–0.99, group-level target JS is 0.03–0.04, and held-out label AUROC is within 0.01 of the oracle. At $N = 121$, collapse to a near-constant predictor occurs in 1/5 replicates for $K = 53$, 3/5 for $K = 256$, and 4/5 for $K = 1024$. At $K = 1024$ and $N = 2000$, the default procedure also collapses in 4/5 replicates.

Eight random restarts and a gradient tolerance tightened by a factor of 10^7 reproduce the same objective in the failing cases, supporting an explanation in terms of collapsed stationary solutions rather than early stopping. Near $w = 0$, core mass equals the query's fraction of core paths, whose variance scales as $1/K$, allowing the readout to absorb the labels. Initialization along the first-order direction $\sum_v (y_v - \bar{y})\,(\bar{\phi}_{\text{core},v} - \bar{\phi}_{\neg\text{core},v})$ reduces failures from 4/5 to 1/5 for $K = 1024$, $N = 2000$, and improves JS from 0.28 to 0.08, but does not resolve the $N = 121$ regime. Collapse is detectable without a planted reference by comparing training loss with a constant predictor.

The fitted real-graph scorers are reproducible under the reported checks. The frozen AlzKB scorer has objective 0.2428 and weight cosine 1.000 across the tested restarts and tolerances. Hetionet reaches objective 0.238 versus 0.640 for the constant predictor, with cosine 1.000 between tested initializations. These diagnostics distinguish recovery behavior in the synthetic regimes from the reproducibility of the two reported real-graph fits.

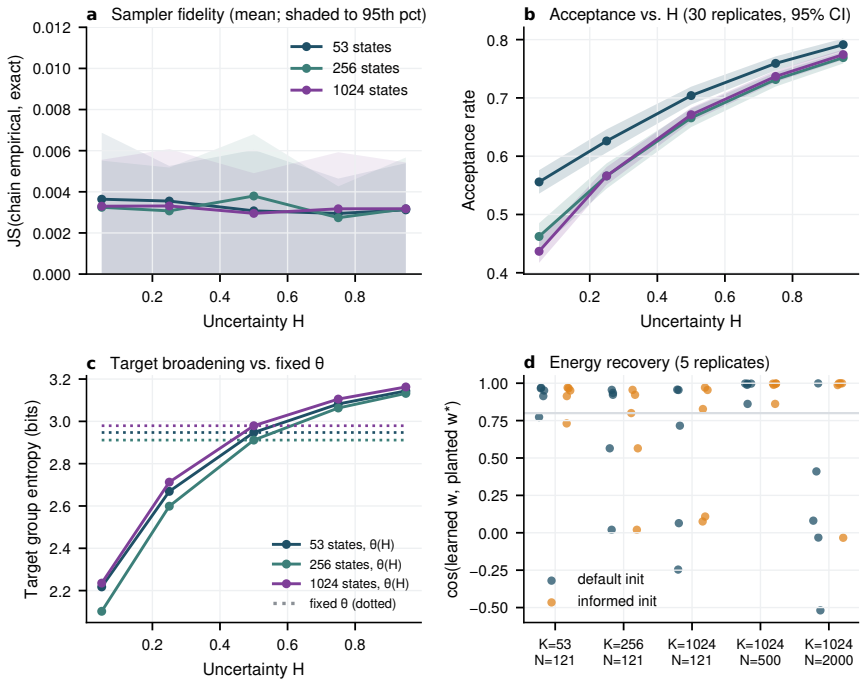


Figure 6: Synthetic checks. (a) Sampler JS versus H for three state-space sizes (means; shading to the 95th percentile; axis from zero; legend top right, shared by (a)–(c)). (b) Acceptance versus H with 95% intervals over 30 replicates. (c) Target group entropy under $\theta(H)$ (solid) and fixed θ (dotted; legend bottom right). (d) Energy recovery: cosine between learned and planted weights per replicate for default (navy) and informed (orange) initialization (legend bottom left); the grey line at 0.8 separates collapsed fits.

E PERTURBATION SENSITIVITY

Perturbing H by ± 0.30 at fixed p produces lower distributional JS at higher uncertainty: the stratum mean decreases from 0.0353 to 0.0074 ($\rho = -0.677$, gene-cluster 95% CI -0.818 to -0.276). The top-1 switch rate increases from 0.004 to 0.150 ($\rho = 0.351$, 95% CI 0.006–0.467). These measurements distinguish two forms of sensitivity: when probability mass is broadly distributed, small changes can exchange the highest-ranked phenotype while leaving the overall hypothesis distribution nearly unchanged.

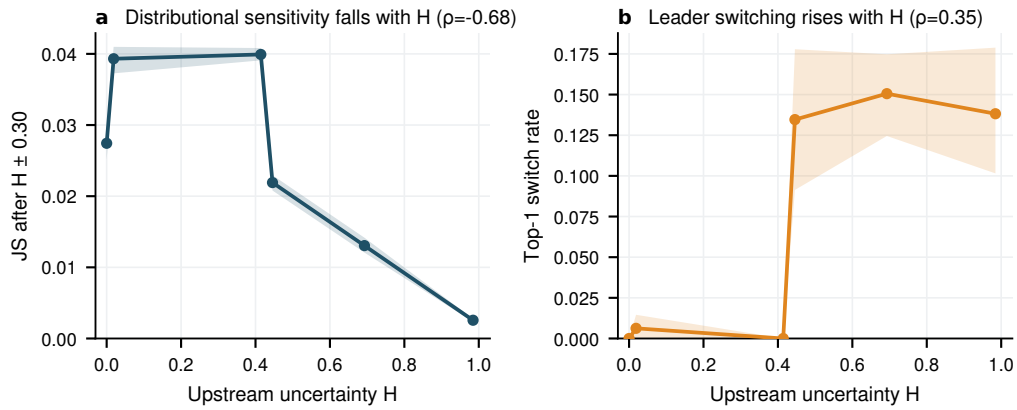


Figure 7: Perturbation of H by ± 0.30 at fixed p (AlzKB). (a) Distributional JS falls with H. (b) Top-1 switch rate rises with H (switches are averages of two binary outcomes, hence the discrete values). Quantile-binned means with 95% bootstrap bands.

F MASKED-EDGE RECOVERY: PROTOCOL CHECKS AND COVERAGE

Leakage audit (seed 42): of the 671 masked disease–symptom edges, 0 remain in the reduced graph in the forward direction, 0 in the inverse direction, 0 appear in any enumerated candidate path, and the relation count used for features equals $3{,}357 - 671$; path and mechanism caches are cleared before enumeration. Coverage, and how it is weighted. Masking removes 601 (gene, symptom) pairs from the 15 evaluated genes—not 601 distinct symptoms, since a symptom masked for two genes is counted once per gene (280 distinct symptoms are masked graph-wide). Of those pairs, 157 remain reachable through gene relays and enter the candidate lists: a pair-weighted reachability of 26%. Recall is computed per query and averaged over the 964 queries with equal weight, so reported values are conditional on the reachable positives of that query's gene, recall@10 $= \frac{1}{964}\sum_v r_{10}(v) = 0.163$. The unconditional value multiplies each query's recall by its own gene's reachable fraction $\kappa_{g(v)}$ before averaging, $\frac{1}{964}\sum_v r_{10}(v)\,\kappa_{g(v)} = 0.100$ (0.237 at $k = 25$). The two do not stand in the ratio 0.26 because reachability is strongly gene-dependent (κ from 0.03 for NOS3 to 0.80 for APP) and the query set is dominated by three high-reachability genes: the query-weighted mean of κ is 0.65, against 0.26 pair-weighted and 0.43 gene-balanced. Gene-held-out splits (Table 5) keep 54–119 training anchors and both calibration classes in every split.

Table 5: Gene-held-out splits (identical for the three mask seeds).

Fold	Held-out genes	anchors	calib. P	calib. B
F1	APP	88	66	22
F2	PSEN1	54	21	33
F3	PSEN2	102	74	28
F4	remaining 15 genes	119	79	40

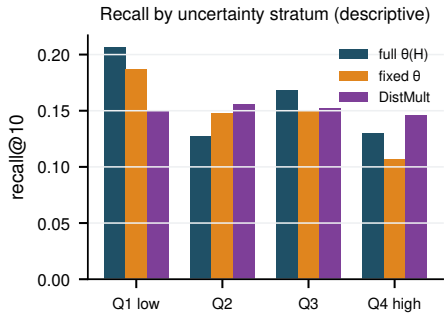
Single-split noninferiority analysis. The primary masked-graph comparison uses a prespecified noninferiority margin of 0.02. Although full-model recall@10 is 0.163 versus 0.151 for DistMult, the paired interval has lower bound -0.044, so this individual-split analysis does not meet the prespecified noninferiority criterion. This is separate from the positive pooled gene-held-out comparison across three masks reported in Section 5.7.

Popularity and per-gene comparisons. On the primary split, symptom popularity (degree in the reduced graph) has mean recall@10 of 0.118, with substantial variation across genes (gene-cluster 95% CI 0.011–0.481). The full-minus-popularity difference is 0.045 (95% CI -0.137 to 0.171). In the gene-held-out comparison, the proposed method has higher recall than DistMult for 8 of the 15 evaluated genes; the advantage is measured in aggregate, not uniformly for every gene. Gene-

balanced full-model recall@10 ranges from 0.237 to 0.312 across masks, and the full-minus-chance difference under gene hold-out is 0.141 (95% CI 0.117–0.146). These summaries accompany, rather than replace, the query-weighted results of Table 2.

G RECALL BY UNCERTAINTY STRATUM (DESCRIPTIVE)

On the primary hold-out, full-model recall@10 is 0.207 in the lowest-H quartile and 0.130 in the highest. This stratified summary characterizes how recovery varies with the upstream signal; the matched-control and gene-held-out experiments of Section 5.7 provide the comparative recovery evidence.



H MASKED-EDGE RECOVERY: A WORKED QUERY

Figure 8 shows one query of the Hetionet masked-edge test. The edge Alzheimer's disease→Aphasia is among the 671 hidden edges, so the query (an APP variant of uncertain significance, $p = 0.95$, $H = 0.44$) never sees it. Aphasia is still reachable on G' through three gene relays (GJA1→autistic disorder, HES5→brain cancer, JUN→epilepsy syndrome); its endpoint mass is the sum of those three path probabilities. After filtering the 70 symptoms still attached to APP's diseases (Alzheimer's disease, lung cancer), 215 candidates remain; Aphasia ranks second under the full model and under fixed θ, and 42nd under DistMult. Of the query's 12 recoverable held-out symptoms, 2 fall in the top 10 and 5 in the top 25; across the 88 (gene, held-out symptom) pairs at the per-gene median-H VUS, 19 rank in the top 10. The pair was chosen for legibility (an Alzheimer's gene and an Alzheimer's symptom with several relays), not as a representative case; the aggregate numbers are those of Table 3. Every element of the figure is read from the experiment's output tables.

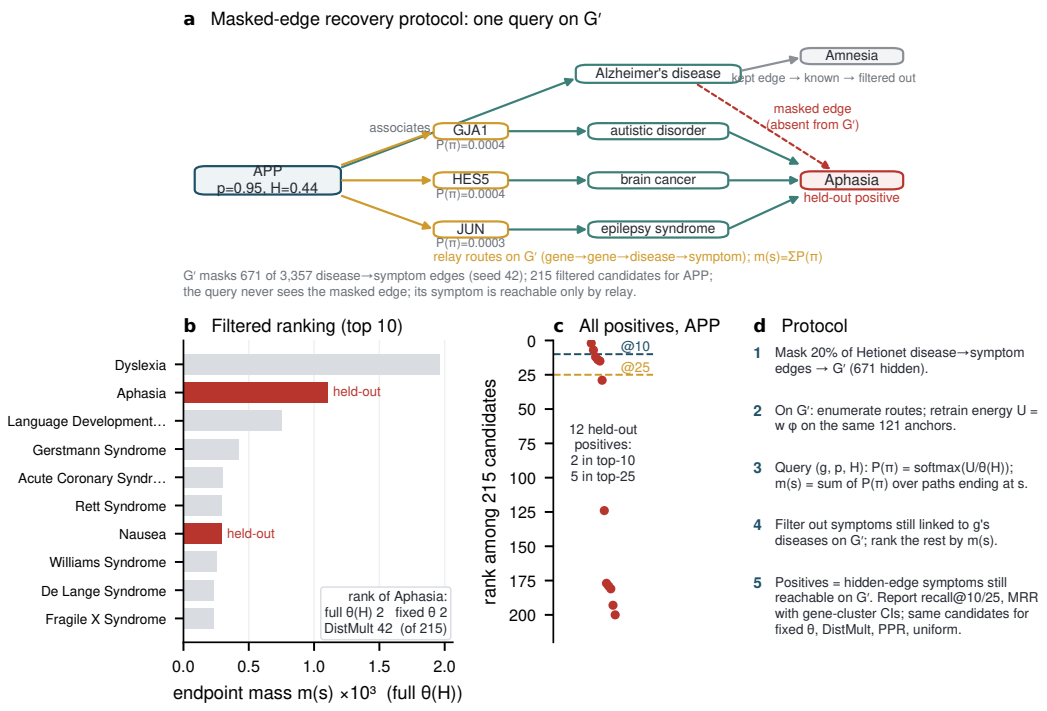


Figure 8: Worked masked-edge recovery query (Hetionet G'). (a) The hidden edge (dashed) and the three relay routes through which the held-out symptom remains reachable; a kept edge of the same disease is filtered from the candidate list. (b) Top 10 of the 215 filtered candidates by endpoint mass under the full model; held-out positives in red, with the target's rank under fixed θ and DistMult. (c) Ranks of all 12 held-out positives of this query. (d) The protocol.

I VARIANT-CLASSIFICATION DIAGNOSTIC IN FULL

The classification diagnostic evaluates variant labels separately from phenotype-hypothesis recovery. Sequence calibration, energy learning, and classification use 121 definitive variants; likely-label records remain locked, and five outer folds refit the upstream pipeline. On 398 valid locked examples, AUROC is 0.955 for sequence-only features and 0.968 after adding graph features. The paired difference is 0.0126 (gene-cluster 95% CI −0.0235 to 0.0191). Brier score changes from 0.112 to 0.120 and expected calibration error from 0.193 to 0.219. In the single-split baseline check, PageRank and uniform features add zero AUROC, while beam features add 0.0085 (95% CI −0.038 to 0.019). These classification and calibration measurements are distinct from the masked-edge recovery task; outputs for 463 valid variants of uncertain significance remain exploratory scores.

J REPRODUCIBILITY AND SPLIT FIREWALL

The corrected suite stores definitive, likely, and VUS IDs in an immutable split manifest. The learned scorer records its training IDs. Five outer-fold manifests verify disjoint training and validation records. Raw input, model, table, figure, and manuscript hashes are frozen. Eight tests cover path validity, feature dimension, detailed balance, empirical stationarity, full proposal support, empty/singleton supports, deterministic seeds, and mass conservation. The mixing experiment reproduces the paper's protocol as a self-test (mean JS 0.0033 versus 0.0031 under independent seeds).

K ADDITIONAL CASE-STUDY DETAILS

Likely-label and VUS records never fit the sequence calibrators, path energy, or classifier. Confidence intervals for the locked diagnostic resample genes as clusters. Gene-cluster intervals are wider than intervals from independent variant resampling; for the classification diagnostic, the graph-augmented AUROC point estimate is higher and the paired interval includes zero (App. I).