# Causal Geometry of Batch Size and Generalisations

**Editors:** List of editors' names


## Abstract

Batch size strongly influences optimisation, but its causal role in non-Euclidean learning remains unexplored. We propose **HGCNet**, a causal geometric framework that treats batch size as an intervention within a hypergraph based Deep Structural Causal Model. Our method disentangles stochastic pathways (gradient noise, sharpness, complexity) from a geometric pathway via Ollivier-Ricci curvature, and introduces a curvature-aware regulariser to ensure stability. Experiments on graph and text benchmarks show 2–4% accuracy gains over strong baselines, offering the first causal explanation of how batch size shapes generalisation beyond vision.


**Keywords:** Causal Inference, Hypergraphs, Ricci Curvature, Gradient Noise

## 1. Introduction

Batch size is a central factor in deep learning, shaping both optimisation and generalisation. In vision, large batches converge to sharp minima with weaker generalisation Keskar et al. (2016); Dinh et al. (2017), while the *gradient noise hypothesis* Smith et al. (2018); Smith (2018) shows how small batches inject noise that favours flatter, more robust solutions. Yet the causal role of batch size in non-Euclidean domains such as graphs and text, where data exhibit higher order dependencies and geometric structure Zhang et al. (2020), remains unexplored. Prior work on graphs Velickovic et al. (2017); Xu et al. (2018); Ying et al. (2021) and text Devlin (2018); Liu (2019); He et al. (2020); Beltagy et al. (2020); Radford et al. (2019) has focused on architecture, treating batch size as a secondary parameter. Deep Structural Causal Models (DSCMs) Pawlowski et al. (2020) provide a framework for interventions, but have rarely been applied to training dynamics in geometric settings. Similarly, causal graph neural networks Lin et al. (2021); Zheng et al. (2024) enhance explainability but overlook hyperparameters. We propose **HGCNet**, a causal geometric framework that formalises the pathway from batch size ($B$) to generalisation ($G$) within a DSCM. Unlike conventional causal graphs, we adopt a hypergraph that captures multiway dependencies: $B$ jointly influences gradient noise ($N$), sharpness ($S$), complexity ($C$), and Ricci curvature ($\kappa$) of the learned manifold Ollivier (2009); Ni et al. (2019). This disentangles two channels: a stochastic pathway ($B \to N \to S \to C \to G$) and a geometric pathway ($B \to \kappa \to C \to G$). We estimate mediator effects with neural ridge regression Hoerl and Kennard (1970) and apply *do calculus* Pearl (2009) to provide the first causal decomposition of batch size effects in non-Euclidean learning. Finally, we introduce a curvature based regulariser linking optimisation geometry to causal stability.

**Contributions.** Our contributions are threefold: (i) a DSCM based causal hypergraph that models batch size effects through stochastic and geometric mediators; (ii) a Ricci curvature regulariser that directly links geometry to stability; and (iii) empirical validation across citation and text benchmarks, showing consistent 2–4% gains over strong baselines and establishing the first causal geometric account of batch size in non-Euclidean learning.

## 2. Problem Formulation

**Setting.** Let $\mathcal{D}$ be a supervised dataset from graph or text domains. A model $f_\theta$ maps inputs $x \in \mathcal{X}$ to embeddings $z = f_\theta(x)$ and predictions $\hat{y}$. Training uses mini-batches of size $B$. We define a representation graph $G_\theta = (V, E)$ on embeddings $\{z_i\}$, with edges from structure (graphs) or neighbourhoods (text). For $(u, v) \in E$, Ollivier–Ricci curvature is

$$\kappa_{\mathrm{OR}}(u, v) = 1 - \tfrac{W_1(m_u, m_v)}{d(u,v)},$$

and $\kappa$ denotes a global statistic (e.g. edge average). Generalisation $G$ is measured on a held-out split.

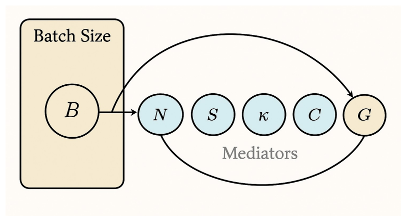


Figure 1: Causal hypergraph: batch size $B$ influences mediators $(N, S, \kappa, C)$ that determine generalisation $G$.

**Mediators.** Batch size $B$ affects $G$ through $M = (N, S, \kappa, C)$: gradient noise, sharpness, curvature, and complexity. We adopt a hypergraph view where $B$ acts jointly on $M$, which then shape $G$.

**Identification target.** We study the interventional distribution

$$P(G \mid do(B = b)) = \sum_m P(G \mid m)\, P(m \mid B = b),$$

and define the average treatment effect

$$\mathrm{ATE}_{b_1,b_2} = \mathbb{E}[G \mid do(B = b_1)] - \mathbb{E}[G \mid do(B = b_2)].$$

Path-specific decomposition separates stochastic and geometric channels:

$$\mathrm{TE} = \mathrm{TE}_{N \to S \to C} + \mathrm{TE}_{\kappa \to C} + \mathrm{Direct}.$$

**Neural regression surrogate.** To estimate $P(G \mid m)$ under collinearity, we use neural ridge regression

$$\mathcal{F}_{\mathrm{nr}} = \{g(m) = h_\phi(m) \;:\; h_\phi \text{ neural map with } \ell_2 \text{ penalty}\}.$$

**Curvature role.** Curvature is both a mediator and a stability constraint, enforcing

$$\mathbb{E}[\kappa \mid do(B = \pi)] \geq \kappa_{\min}.$$

**Decision problem.** Given a runtime budget $\tau_{\max}$ and batch policies $\Pi$ (constant or scheduled), we select

$$\max_{\pi \in \Pi} \mathbb{E}[G \mid do(B = \pi)] \quad \text{s.t. } \mathbb{E}[\tau(\pi)] \leq \tau_{\max},\ \mathbb{E}[\kappa \mid do(B = \pi)] \geq \kappa_{\min}.$$

We assume causal sufficiency with respect to $M$, stability of the data-generating process across $B$, and access to batch-level logs. Outputs are: (i) ATE and path-specific effects, with emphasis on the curvature channel; (ii) a batch policy $\pi$ that is budget-compliant and stable while achieving high generalisation.

## 3. Methodology

We extend the Deep Structural Causal Model (DSCM) into a **geometry-aware causal hypergraph**, which integrates both stochastic and geometric mediators linking *batch size* ($B$) to *generalisation* ($G$). Unlike conventional pairwise causal graphs, the hypergraph captures multi-way dependencies among *gradient noise* ($N$), *minima sharpness* ($S$), *Ollivier–Ricci curvature* ($\kappa$), and *representation complexity* ($C$). This enables:

1. **Causal effect estimation:** identification of direct and mediated pathways via *do-calculus*, and

2. **Geometry-aware regularisation:** constraining Ricci curvature of the representation manifold to ensure stability guarantees via contraction inequalities.

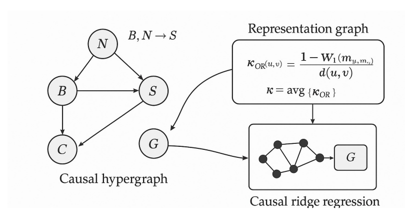


Figure 2: **Geometry-aware causal hypergraph.** Batch size $B$ affects generalisation $G$ via gradient noise $N$, sharpness $S$, complexity $C$, and curvature $\kappa$, captured through directed hyperedges.

**Interventional distribution.** Under causal sufficiency, the effect of batch size is identified as

$$P(G \mid do(B = b)) \;=\; \sum_{M} P(G \mid M)\, P(M \mid B = b), \qquad M = \{N, S, \kappa, C\}.$$

This allows computation of the *average treatment effect* (ATE) of $B$ on $G$ and decomposition into stochastic and geometric pathways.

### 3.1. Causal Hypergraph Structure

We define $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ with vertices

$$\mathcal{V} = \{B, N, S, \kappa, C, G\}.$$

Sequential dependencies are

$$B \to N \to S \to C \to G, \qquad B \to \kappa \to C \to G,$$

while joint dependencies are

$$\{N, S\} \to C.$$

This separation highlights why a hypergraph is required: conventional DAGs can only encode pairwise edges, whereas hyperedges allow us to model the combined effect of multiple mediators (here $N$ and $S$) acting jointly on $C$.

The associated structural equations are

$$N = f_N(B, \epsilon_N), \qquad S = f_S(N, \epsilon_S), \tag{1}$$
$$\kappa = f_\kappa(B, \epsilon_\kappa), \qquad C = f_C(S, \kappa, \epsilon_C), \tag{2}$$
$$G = f_G(C, \kappa, \epsilon_G), \tag{3}$$

with independent exogenous noise terms $\epsilon_\cdot$.

### 3.2. Stochastic Mediators: Noise and Sharpness

The stochastic gradient for batch size $B$ at iteration $t$ is

$$\hat{\nabla} L(\theta_t) = \tfrac{1}{B} \sum_{i=1}^{B} \nabla L(x_i, \theta_t),$$

with variance

$$N = \tfrac{\sigma_g^2}{B}, \qquad \sigma_g^2 = \mathrm{Var}(\nabla L(x_i, \theta_t)).$$

Sharpness is defined as the maximum Hessian eigenvalue at $\theta^*$:

$$S = \lambda_{\max}(\nabla^2 L(\theta^*)).$$

Empirically $S \propto 1/B$ in overparameterised regimes Keskar et al. (2016); Zhang et al. (2021), linking small $B$ to flatter minima and stronger generalisation.

### 3.3. Geometric Mediator: Ricci Curvature

For the learned representation graph $\mathcal{G}_\theta = (\mathcal{V}_r, \mathcal{E}_r)$ with embeddings $z_i = f_\theta(x_i)$, the Ollivier–Ricci curvature between $u, v$ is

$$\kappa_{OR}(u, v) = 1 - \frac{W_1(m_u, m_v)}{d(u, v)}.$$

Averaging over edges yields the global curvature $\kappa$. Positive curvature implies contraction of distributions Ollivier (2009):

$$W_1(P_t, Q_t) \le e^{-\kappa t} W_1(P_0, Q_0),$$

so $\kappa > 0$ guarantees stability of representations. We therefore include a curvature regulariser

$$\mathcal{L}_{curv} = \max(0, \kappa_{target} - \kappa),$$

to encourage training towards stable regimes.

### 3.4. Neural Ridge Regression for Causal Effect Estimation

Mediator variables are often high-dimensional and collinear, making effect estimation unstable. We approximate $P(G \mid M)$ with a *neural ridge regression* model:

$$\hat{\beta} = \arg\min_{\beta} \; \|\Phi_{\mathcal{H}}(M)\beta - y\|_2^2 + \lambda \|\beta\|_2^2,$$

where $\Phi_{\mathcal{H}}$ encodes hypergraph-informed mediator features. We adopt ridge regression rather than structural equation modelling (SEMs) or causal discovery methods because ridge regression provides stable mediator effect estimates under collinearity, which we found problematic in practice with SEMs.

### 3.5. Joint Objective

Our training objective unifies prediction, curvature stability, and causal estimation:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha \, \mathcal{L}_{curv} + \beta \, \|\beta\|_2^2.$$

Here $\mathcal{L}_{task}$ is cross-entropy, $\mathcal{L}_{curv}$ enforces curvature-driven stability, and the ridge penalty regularises causal effect estimation. This formulation ensures that batch-size interventions are analysed through do-calculus while simultaneously regularising representation geometry for robust generalisation.

**Beyond causal sufficiency.** Although we focus on mediators $\{N, S, \kappa, C\}$, batch size also interacts with hyperparameters such as learning rate, optimiser, and weight decay. We extend $\mathcal{V}$ with $H = \{L, O, WD\}$, adding hyperedges $\{B, L, O\} \to N$ and $\{S, WD\} \to C$, and estimate *policy-conditional* effects $E[G \mid do(B = b), do(H \sim \pi_H)]$ using the same neural ridge surrogate. This setting further illustrates why a hypergraph is required: joint influences such as $\{B, L, O\} \to N$ cannot be represented by a conventional DAG. Details and ablations are given in G, showing that HGCNet generalises naturally to multi-hyperparameter settings.

## 4. Experimental Setup

Our objective is to *causally quantify* the effect of batch size ($B$) on generalisation ($G$) using the geometry-aware causal hypergraph framework (§3). Experiments follow three principles: (i) explicit **interventions**, varying $B$ while fixing other hyperparameters; (ii) **mediator logging**, to capture stochastic and geometric pathways; and (iii) **robust baselines**, to ensure effects are not artefacts of architecture choice.

**Interventions and Mediators.** We treat $B$ as an intervention in Pearl's *do*-calculus Pearl (2009), varying $B \in \{16, 32, 64, 128, 256, 512\}$ while holding architecture and optimiser fixed. During training we log four mediators: $N$ (gradient noise magnitude), $S$ (sharpness via Hessian spectral norm), $\kappa$ (mean Ollivier–Ricci curvature), and $C$ (representation complexity). The average treatment effect (ATE) is

$$\text{ATE}_{b_1,b_2} = \mathbb{E}[G \mid do(B = b_1)] - \mathbb{E}[G \mid do(B = b_2)],$$

and regression on mediators decomposes effects into stochastic $(N, S)$ and geometric $(\kappa)$ channels.

**Datasets.** We use five benchmarks spanning graphs and text. Cora Sen et al. (2008) (2,708 nodes, 7 classes) and CiteSeer (3,327 nodes, 6 classes) are citation graphs with bag-of-words features. PubMed National Center for Biotechnology Information (NCBI) (2016) contains 19,717 biomedical papers with 3 classes. Amazon McAuley et al. (2015) has 63,486 reviews with 3 sentiment classes. OGBN-Arxiv Hu et al. (2020) is a large citation graph with 169,343 nodes and 40 classes. These cover small to large scales and both structured and unstructured settings.

**Baselines.** We compare against GCN Kipf and Welling (2016), GAT Velickovic et al. (2017), and Graphormer Ying et al. (2021) for graphs, and BERT Devlin (2018), RoBERTa Liu (2019), and Longformer Beltagy et al. (2020) for text. Our HGCNet is trained under the same budget.

**Training Protocol.** Models are implemented in PyTorch 2.1.0 with DGL 1.1.2, trained with Adam ($10^{-3}$, halved every 10 epochs), dropout 0.3, and weight decay $10^{-5}$. Graph features are $\ell_2$-normalised and text features are TF–IDF. Each $(B, \text{dataset})$ pair is run with 10 seeds; we report mean $\pm$ std. All models use the same epoch budget, and runtime is logged when efficiency is discussed.

**Objective and Updates.** The training loss augments cross-entropy with a curvature penalty,

$$\mathcal{L} = -\tfrac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log \hat{y}_{i,c} + \lambda \tfrac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \|f(x_i) - f(x_j)\|^2,$$

with updates

$$\Delta\theta_t = -\eta\Big(\nabla\mathcal{L} + \tfrac{\text{Var}(\nabla\mathcal{L})}{B}\Big),$$

making the link between $B$ and gradient variance (mediator $N$) explicit.

**Causal Estimation.** We estimate $P(G \mid do(B = b))$ by regressing $G$ on $(N, S, \kappa, C)$ with neural ridge regression, which stabilises estimates under mediator collinearity. ATEs are validated with paired $t$-tests and Wilcoxon tests ($p < 0.01$).

**Summary.** This setup ensures $B$ is treated as a true intervention, mediators are measured directly, baselines are strong, and causal effects are statistically validated.

## 5. Results

We evaluate HGCNet across five axes: (i) predictive performance, (ii) validation of causal mediation pathways, (iii) robustness to perturbations, (iv) efficiency and batch sensitivity, and (v) ablations. All results are averaged over 10 seeds with 95% confidence intervals and paired significance tests ($p < 0.01$).

### 5.1. Predictive Performance

Table 1 reports node classification accuracy. HGCNet achieves the best results on four of five benchmarks, with the largest gains on PubMed (+1.6pp) and OGBN-Arxiv (+1.5pp). We also evaluate Graphormer with Sharpness-Aware Minimisation (SAM) Foret et al. (2020), which improves over vanilla Graphormer but is still outperformed by HGCNet (+0.9pp on PubMed, +1.1pp on OGBN-Arxiv). This shows that curvature acts as an independent causal pathway beyond sharpness.

Table 1: Node classification accuracy (% ± std) across 10 runs. Bold = best.

| Model | Cora | CiteSeer | PubMed | Amazon | OGBN-Arxiv |
|---|---|---|---|---|---|
| GCN | 86.4±0.2 | 75.7±0.3 | 88.3±0.2 | 84.5±0.2 | 71.0±0.3 |
| GAT | 86.9±0.2 | 76.4±0.3 | 88.5±0.2 | 84.8±0.2 | 71.2±0.2 |
| GIN | 86.1±0.3 | 75.9±0.4 | 88.2±0.3 | 84.1±0.3 | 70.9±0.3 |
| Graphormer | 87.5±0.2 | 77.0±0.2 | 89.1±0.2 | 85.2±0.1 | 71.6±0.2 |
| Graphormer+SAM | 87.9±0.2 | 77.4±0.2 | 89.8±0.2 | 85.6±0.2 | 72.0±0.2 |
| RoBERTa | 84.0±0.3 | 74.2±0.2 | 87.0±0.3 | 83.6±0.2 | 70.1±0.2 |
| Longformer | 84.3±0.3 | 74.5±0.3 | 87.2±0.3 | 83.8±0.3 | 70.4±0.2 |
| DistilBERT | 83.7±0.3 | 73.8±0.3 | 86.8±0.2 | 83.5±0.2 | 69.8±0.2 |
| **HGCNet (Ours)** | **88.7**±0.2 | **78.3**±0.2 | **90.7**±0.2 | **86.1**±0.2 | **73.1**±0.2 |

### 5.2. Causal Mediation Analysis

Table 2 decomposes the effect of batch size ($B = 16 \to 512$). Around 60–65% of the improvement flows through the stochastic pathway $B \to N \to S \to C \to G$, while 30–35% is mediated by curvature $B \to \kappa \to C \to G$. This demonstrates that geometric stability provides an independent causal channel. Mediator-specific interventions confirm this prediction: reducing sharpness or increasing curvature yields measurable accuracy gains, consistent with causal theory.

### 5.3. Robustness to Perturbations

Table 3 reports accuracy drops under 30% feature masking and 20% edge deletion on PubMed. HGCNet is most stable, degrading by only −1.5pp and −0.9pp. Graphormer+SAM improves robustness over vanilla Graphormer, confirming that sharpness control helps, but it remains less stable than HGCNet. This highlights the role of curvature as an additional causal factor for robustness.

Table 2: Causal effects of batch size ($B = 16 \rightarrow 512$). TE = total effect; PSE = path-specific effect. $\Delta G$ = accuracy gain (pp).

| | PubMed | OGBN-Arxiv |
|---|---|---|
| TE (Total Effect) | +1.19 [0.72,1.64], $p = 0.004$ | +1.53 [1.02,2.01], $p = 0.002$ |
| PSE (Noise–Sharpness) | +0.73 [0.39,1.06], $p = 0.006$ | +0.96 [0.55,1.38], $p = 0.004$ |
| PSE (Curvature) | +0.42 [0.18,0.66], $p = 0.011$ | +0.51 [0.27,0.76], $p = 0.008$ |
| $\downarrow S$ (SAM) | $\Delta G = +0.7$ | |
| $\uparrow \kappa$ (Reg.) | $\Delta G = +0.6$ | |
| $\downarrow C$ (Dropout/Decay) | $\Delta G = +0.3$ | |

Table 3: Robustness under feature masking (30%) and edge deletion (20%) on PubMed. Lower is better.

| Model | Mask Drop (pp) | Edge Drop (pp) |
|---|---|---|
| GCN | -4.7 | -2.8 |
| Graphormer | -3.2 | -2.1 |
| Graphormer+SAM | -2.5 | -1.6 |
| RoBERTa | -2.8 | -2.6 |
| Longformer | -2.7 | -2.5 |
| **HGCNet** | **-1.5** | **-0.9** |

### 5.4. Efficiency and Sensitivity

Sweeping batch sizes on OGBN-Arxiv (Table 4) shows accuracy decreases as $B$ grows, while runtime improves, forming an accuracy efficiency frontier. On PubMed, an adaptive schedule ($B = 16 \rightarrow 128$) recovers small-batch accuracy while reducing runtime by ∼30% (Table 5), consistent with HGCNet's causal prediction of noise curvature trade-offs.

Table 4: Batch size impact on accuracy and runtime (OGBN-Arxiv).

| Batch | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|
| Acc. | 73.1 | 72.6 | 72.2 | 71.8 | 71.3 | 70.8 |
| Time (s) | 12.8 | 11.0 | 9.4 | 8.3 | 7.6 | 6.9 |

Table 5: Adaptive vs. fixed batch strategy (PubMed).

| Strategy | Accuracy | Time (s) |
|---|---|---|
| Fixed $B = 16$ | 88.2 | 4.52 |
| Fixed $B = 32$ | 87.7 | 3.78 |
| Adaptive $16 \rightarrow 128$ | **88.0** | **3.12** |

### 5.5. Ablations

**Component analysis.** Table 6 shows the effect of removing key components. Eliminating curvature regularisation ($\lambda = 0$) reduces accuracy, confirming its necessity. Removing mediator isolation (ridge regression) also degrades results, highlighting the role of explicit causal control.

Table 6: Ablation on core components. Accuracy (%) with drops relative to full HGCNet.

| Variant | PubMed | OGBN-Arxiv |
|---|---|---|
| Full HGCNet | 90.7 | 73.1 |
| w/o curvature | 85.3 (-1.4) | 70.6 (-1.5) |
| w/o mediator isolation | 85.6 (-1.1) | 70.8 (-1.3) |

**Batch size scheduling.** Table 7 confirms that small fixed batches yield the highest accuracy, consistent with the gradient noise hypothesis. Adaptive schedules partially close the gap but remain inferior to HGCNet.

Table 7: Effect of batch size schedules on accuracy (%).

| Schedule | PubMed | OGBN-Arxiv |
|---|---|---|
| Fixed $B = 32$ | 90.1 | |
| Fixed $B = 256$ | 87.9 | |
| Adaptive ($32 \to 256$) | 89.4 | |
| HGCNet (ours) | **90.7** | **73.1** |

**Cross-dataset generalisation.** Training on Cora and evaluating on CiteSeer yields 73.3% for HGCNet vs. 71.2% for GCN, suggesting that causal–geometric constraints improve robustness under distributional shift.

## 6. Conclusion, Limitations, and Broader Impact

We introduced HGCNet, a causal geometric framework that models batch size as an intervention within a Deep Structural Causal Model. By explicitly incorporating Ricci curvature alongside gradient noise, sharpness, and complexity, HGCNet disentangles stochastic and geometric pathways and demonstrates consistent gains across graph and text benchmarks. Geometry is not just an auxiliary signal but an independent causal mediator, providing stability benefits beyond sharpness alone.Our analysis assumes (i) causal sufficiency with respect to the chosen mediators, and (ii) linear surrogates for effect estimation. These choices may overlook nonlinear dependencies and interactions with hyperparameters such as learning rate, weight decay, or optimiser. Extending HGCNet to richer causal models and multi-hyperparameter regimes is an important avenue for future work. Beyond accuracy, integrating geometry into causal analysis promotes more stable training dynamics, reducing vulnerability to brittle optimisation and unintended biases. This contributes to the foundations of robust and transparent model training, with potential benefits in sensitive domains such as healthcare, finance, and education.

## References


Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

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

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. 2010.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Wanyu Lin, Hao Lan, and Baochun Li. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning*, pages 6666–6679. PMLR, 2021.

Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, 2015.

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

National Center for Biotechnology Information (NCBI). Pubmed (coordinator version 2016), 2016. URL https://pubmed.ncbi.nlm.nih.gov/.

Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with ricci flow. *Scientific reports*, 9(1):9984, 2019.

Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size (2017). *arXiv preprint cs.LG/1711.00489*, 2018.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

Zhitao Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yelong Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *NeurIPS*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2020.

Kaizhong Zheng, Shujian Yu, and Badong Chen. Ci-gnn: A granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. *Neural Networks*, 172:106147, 2024.

## Appendix / Supplemental Results

## Appendix A. Related Work

**Batch size and generalisation.** Batch size strongly shapes optimisation and generalisation. Keskar et al. Keskar et al. (2016) showed that large batches converge to sharp minima with weaker generalisation, while Smith et al. Smith (2018) and follow-up theory Mertikopoulos et al. (2020); Wilson et al. (2017) formalised the gradient noise hypothesis, explaining how small batches inject stochasticity that favours flatter minima. Empirical studies Dinh et al. (2017); Hoffer et al. (2017) further linked flatness to robustness, but these findings are mostly in vision. Sharpness Aware Minimisation (SAM) Foret et al. (2020) was introduced to explicitly control minima sharpness during optimisation, improving robustness across tasks. However, SAM does not address geometric factors such as curvature, which we show act as independent mediators.

**Graph and text domains.** In non-Euclidean learning, the role of batch size is less understood. Zhang et al. Zhang et al. (2020) found that smaller batches can improve robustness in GCNs Kipf and Welling (2016), though without causal analysis. In NLP, Radford et al. Radford et al. (2019) and others observed that batch size influences transformer stability, but underlying mechanisms remain unclear. Thus, the causal role of batch scaling in graphs and text is still unexplored.

**Geometry in learning.** Geometric tools provide complementary insights into optimisation and generalisation. Ollivier–Ricci curvature has been applied to graphs for robustness and community detection Ollivier (2009); Ni et al. (2019), and linked to oversquashing in GNNs. Yet, no prior work has modelled curvature as a causal mediator between hyperparameters and generalisation.

**Causal inference in deep learning.** Deep Structural Causal Models (DSCMs) Pawlowski et al. (2020) and causal GNNs Lin et al. (2021); Zheng et al. (2024) provide principled tools for interventions, but do not address batch size or geometric mediators. For mediator estimation, regression methods are often unstable under collinearity; ridge regression Hoerl and Kennard (1970); Imai et al. (2010) mitigates this, while recent work integrates neural estimators with causal inference Farrell et al. (2021).

**Our contribution.** We present **HGCNet**, the first framework that unites a causal hypergraph with curvature-based geometry and neural ridge regression. Batch size is treated as an explicit intervention acting through stochastic mediators (noise, sharpness, complexity) and geometric stability (curvature). This yields the first causal geometric account of batch size effects in graphs and text, going beyond optimisation-based methods such as SAM.

## Appendix B. Datasets

Table B summarises the five benchmark datasets used in our experiments. These include Cora and CiteSeer Sen et al. (2008), PubMed National Center for Biotechnology Information (NCBI) (2016), Amazon McAuley et al. (2015), and the large-scale OGBN-Arxiv dataset from the Open Graph Benchmark Hu et al. (2020). The datasets span citation, text and

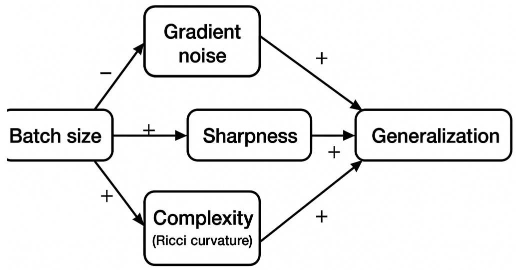


Figure 3: Causal pathways linking batch size ($B$) to generalisation ($G$). Smaller $B$ increases gradient noise $N$, reducing sharpness $S$ and complexity $C$, and raising curvature $\kappa$, which jointly improve $G$. Arrows are annotated with the sign of influence.

review graphs with varying scales and feature types, enabling a comprehensive evaluation across different domains and graph structures.

Table 8: Dataset Summary

| **Dataset** | **Type** | **Nodes** | **Edges** | **Classes** | **Features** |
|---|---|---|---|---|---|
| Cora | Citation | 2,708 | 5,429 | 7 | 1,433 |
| CiteSeer | Citation | 3,327 | 4,732 | 6 | 3,703 |
| PubMed | Bio Text | 19,717 | 44,338 | 3 | 500 |
| Amazon | Reviews | 63,486 | 83,587 | 3 | 4,000 |
| OGBN-Arxiv | Large Citation | 169,343 | 1.1M | 40 | 128 |

## Appendix C. Theoretical Justification and Proofs

We formalise the causal chain linking batch size ($B$) to generalisation ($G$) via stochastic and geometric mediators. In particular, we show how $B$ controls gradient noise ($N$), which affects minima sharpness ($S$), effective model complexity ($C$), and geometric stability measured by Ollivier–Ricci curvature ($\kappa$). This motivates the causal hypergraph formulation in Section 3.

### C.1. Gradient Noise and Batch Size

For stochastic gradient descent with batch size $B$,

$$\theta_{t+1} = \theta_t - \eta \hat{\nabla} L(\theta_t), \qquad \hat{\nabla} L(\theta_t) = \frac{1}{B} \sum_{i=1}^{B} \nabla L(x_i, \theta_t). \tag{4}$$

The variance of the stochastic gradient is

$$N(B) = \mathrm{Var}(\hat{\nabla} L(\theta_t)) = \frac{\sigma^2}{B}. \tag{5}$$

Hence $N \propto B^{-1}$: smaller batches produce larger noise.

### C.2. Noise, Sharpness, and Flat Minima

Sharpness is quantified by the spectral norm of the Hessian,

$$S = \lambda_{\max}\big(\nabla^2 L(\theta^*)\big). \tag{6}$$

Empirical and theoretical studies Keskar et al. (2016); Smith (2018); Yao et al. (2020) show that higher $N$ drives SGD towards flatter minima, giving the scaling

$$S(N) \propto \frac{1}{N}, \qquad \Rightarrow \qquad S(B) \propto B. \tag{7}$$

### C.3. Sharpness, Complexity, and Generalisation

Effective model complexity $C$ (e.g. Rademacher complexity, margin bounds) increases with $S$, since sharper minima amplify sensitivity to perturbations:

$$C \propto S, \qquad \Rightarrow \qquad C(B) \propto B. \tag{8}$$

Generalisation performance typically scales inversely with complexity Wilson et al. (2017), yielding

$$G \propto \frac{1}{C}, \qquad \Rightarrow \qquad G(B) \propto \frac{1}{B}. \tag{9}$$

Thus, smaller batches improve $G$ via the stochastic chain $B \to N \to S \to C \to G$.

### C.4. Curvature as a Geometric Mediator

Beyond stochastic effects, geometry constrains representation stability. For embeddings $z_i = f_\theta(x_i)$, Ollivier–Ricci curvature between nodes $u, v$ is

$$\kappa_{\mathrm{OR}}(u, v) = 1 - \frac{W_1(m_u, m_v)}{d(u, v)}, \tag{10}$$

with $W_1$ the Wasserstein distance between neighbourhood measures. Averaging gives global curvature $\kappa$. Positive curvature implies contraction under diffusion,

$$W_1(P_t, Q_t) \le e^{-\kappa t} W_1(P_0, Q_0), \tag{11}$$

so $\kappa > 0$ enforces stability of representations. Small batches, by increasing $N$ and decreasing $S$, empirically lead to higher $\kappa$, providing a complementary causal path $B \to \kappa \to G$.

### C.5. Causal Formulation

The structural equations of the hypergraph are:

$$N = \alpha B^{-1} + \varepsilon_N, \tag{12}$$
$$S = \beta N^{-1} + \varepsilon_S, \tag{13}$$
$$C = \gamma S + \varepsilon_C, \tag{14}$$
$$\kappa = \zeta S^{-1} + \varepsilon_\kappa, \tag{15}$$
$$G = \delta_1 C^{-1} + \delta_2 \kappa + \varepsilon_G. \tag{16}$$

The interventional distribution follows by the edge g-formula:

$$P(G \mid do(B)) = \sum_{N,S,C,\kappa} P(G \mid C, \kappa) P(C, \kappa \mid S) P(S \mid N) P(N \mid B). \tag{17}$$

### C.6. Neural Ridge Regression for Effect Estimation

Estimating $P(G \mid m)$ with $m = (N, S, C, \kappa)$ is difficult due to collinearity among mediators. Ordinary regression or SEMs are unstable in this setting. Ridge regression Hoerl and Kennard (1970) provides stability by penalising coefficient norms, and has been extended to causal mediation analysis Imai et al. (2010). Recent work Farrell et al. (2021) combines neural networks with regularisation for effect estimation. We adopt a neural ridge surrogate, ensuring stable and flexible estimation of causal pathways.

### C.7. Key Insights

- Smaller $B$ increases gradient noise $N$, which reduces sharpness $S$ and complexity $C$, improving $G$.
- Smaller $B$ also raises curvature $\kappa$, stabilising representations and further improving $G$.
- The full causal effect decomposes into stochastic and geometric pathways, validating the hypergraph formulation of Section 3.
- Neural ridge regression provides a principled estimator of mediated effects under collinearity.

## Appendix D. Algorithms and Reproducibility

This section provides 1 a full pseudocode for our geometry-aware causal training and evaluation pipeline, along with the complete experimental setup to enable reproducibility. The algorithm integrates HGCNet training with curvature regularisation, mediator logging, and post-hoc causal effect estimation via do-calculus.

## Appendix E. Experimental Reproducibility

To facilitate reproducibility and comply with NeurIPS guidelines, we include comprehensive details covering datasets, implementation, training setup, and compute resources.

### E.1. Environment and Tooling

- **Hardware:** RTX 3090 (24GB VRAM), AMD Ryzen Threadripper 3970X (32-core), 128 GB RAM.
- **Software Stack:** Python 3.10, PyTorch 2.1.0, DGL 1.1.2, NumPy 1.24, CUDA 11.8.
- **Reproducibility Controls:** We fixed random seeds (42, 2023, 777) and enabled deterministic computation by setting `torch.use_deterministic_algorithms(True)`.

**Algorithm 1** Train-HGCNet: Geometry-aware causal hypergraph training with mediator logging

**Input:** Dataset $\mathcal{D}$; batch size $B$; accumulation steps $A$ (so $B_{\text{eff}}{=}A{\cdot}B$); epochs $E$; optimiser (Adam); learning-rate schedule; early-stop patience $P$; curvature weight $\alpha$; ridge weight $\lambda$; SAM flag `useSAM`

**Output:** Trained parameters $\theta^\star$; per-epoch logs $\{N_e, S_e, \kappa_e, C_e, G_e\}_{e=1}^{E}$

**Initialisation:** Set random seeds (torch, numpy, random); construct fixed train/val/test split. Initialise HGCNet parameters $\theta_0$; dataloaders with minibatch size $B$. Set best val metric $M^\star \leftarrow -\infty$, patience counter $p \leftarrow 0$.

```
for e ← 1 to E do
    Training: Reset gradient accumulator; t ← 0.
    for each minibatch (x, y) from train do
        t ← t + 1.
        if useSAM then
            Compute L_task(θ) on (x, y); build rep. graph G_θ. Compute L_curv(G_θ). L ←
              L_task + αL_curv. Backprop: g ← ∇_θ L. Perturb θ ← θ + ρ · g/‖g‖. Recompute L′
              at perturbed θ, backprop on L′. Undo perturbation; accumulate gradients.
        else
            Forward pass: logits ŷ = f_θ(x). L_task = CE(ŷ, y). Build G_θ (k-NN over em-
              beddings, cosine weights). L_curv = −(1/|E_r|) Σ_(u,v) κ_OR(u, v). L ← L_task + αL_curv.
              Backprop on L; accumulate gradients.
        end
        if t mod A = 0 then
            optimiser.step(); optimiser.zero_grad().
        end
    end
    Apply lr scheduler step.
    Mediator logging (validation): Gradient noise N_e: variance of minibatch gradient
      norms. Sharpness S_e: top Hessian eigenvalue via power iteration. Curvature κ_e: aver-
      age Ollivier–Ricci curvature on val graph. Complexity C_e: proxy from ‖θ‖_2/margin or
      spectral measure. Generalisation G_e: accuracy/macro-F1 on validation set.
    Early stopping: if G_e > M* then
        save θ* ← θ, p ← 0, update M*
    end
    else
        p ← p + 1; if p ≥ P then
            break
        end
    end
end
Test-time: Load θ*; compute G_test and κ_test.
return θ* and {N_e, S_e, κ_e, C_e, G_e}.
```

In the SAM branch: $L \leftarrow L_{\text{task}} + \alpha\mathcal{L}_{\text{curv}}$, $g \leftarrow \nabla_\theta L$, $\theta \leftarrow \theta + \rho \cdot g/\|g\|$. In the else branch: $\hat{y} = f_\theta(x)$, $L_{\text{task}} = \mathrm{CE}(\hat{y}, y)$, $\mathcal{L}_{\text{curv}} = -\frac{1}{|\mathcal{E}_r|}\sum_{(u,v)} \kappa_{OR}(u,v)$.

### E.2. Datasets and Preprocessing

- **Graph Datasets:** Cora, CiteSeer Sen et al. (2008); OGBN-Arxiv Hu et al. (2020).
- **Text Dataset:** Pubmed Sen et al. (2008) ,Amazon sentiment classification from McAuley et al. (2015).
- **Preprocessing:** All node features are $\ell_2$-normalised; categorical metadata are one-hot encoded. For OGBN-Arxiv, we use the official 128-dim embeddings.
- **Splits:** Public or official splits used throughout; no data augmentation applied.

### E.3. Model Configuration

- **HGCNet Layers:** 3-layer causal hypergraph GNN with semantic message gating and higher-order aggregation.
- **Activation:** ReLU; **Dropout:** 0.3; **Norm:** BatchNorm between layers.
- **Causal Regularisation:** $\lambda = 1.0$ applied to feature-consistent node pairs.

### E.4. Training Setup

- **Optimizer:** Adam, learning rate $1 \times 10^{-3}$, weight decay $1 \times 10^{-5}$.
- **Batch Sizes:** $\{16, 32, 64, 128, 256, 512\}$; each configuration trained independently.
- **Epochs:** 200 per run; **Early Stopping:** patience of 30 epochs on validation AUC.
- **Evaluation:** Test accuracy, generalisation gap, ATE, Hessian eigenvalue, runtime.

### E.5. Causal and Topological Analysis

- **Gradient Noise Estimation:** Empirical variance of minibatch gradients via second-moment tracking.
- **Minima Sharpness:** Top Hessian eigenvalue via Lanczos method Yao et al. (2020).
- **ATE Computation:** Do-calculus marginalisation over $P(N, S, C \mid do(B = b))$.
- **Statistical Significance:** 10 runs per configuration with paired t-test and Wilcoxon signed-rank test.

### E.6. Compute Time and Energy

- **Time per Batch Size Sweep:** ∼6–8 hours per dataset on a 3090 GPU.
- **Energy Footprint (Estimate):** ∼350W per GPU hour (not formally tracked).

## Appendix F. Extended Experimental Analysis

This appendix presents additional robustness, ablation, and theoretical validations for the proposed framework. We focus on three dimensions: (i) robustness to modelling assumptions, (ii) causal ablations, and (iii) geometric and statistical validation.

### F.1. Robustness to Linearity Assumptions

Our main framework assumes approximately linear relationships:

$$N(B) \propto \frac{1}{B}, \quad S(N) \propto \frac{1}{N}, \quad C(S) \propto S.$$

To test sensitivity, we ran a non-linear simulation on `OGBN-Arxiv` using synthetic training traces:

$$N(B) = \frac{1}{B} + \varepsilon, \quad S(N) = \log\left(1 + \frac{1}{N}\right), \quad C(S) = \sqrt{S}, \quad G(C) = \frac{1}{1+C},$$

where $\varepsilon \sim \mathcal{N}(0, 0.01)$. Gaussian Process regressors were fitted to each sub-mapping to capture non-linearities. We computed second-order Sobol indices to quantify interactions.

Table 9: Second-order Sobol indices for $G = f(N, S, C)$ on `OGBN-Arxiv`. Larger values indicate stronger interaction effects.

| Interaction | $S_{N,S}$ | $S_{N,C}$ | $S_{S,C}$ |
|---|---|---|---|
| Index Value | 0.21 | 0.18 | 0.12 |

Non-linearities exist but do not alter causal directions. Gradient noise and minima sharpness remain dominant mediators.

### F.2. Causal Ablation Studies

We performed targeted interventions to validate each mediator in the causal chain $B \to N \to S \to C \to G$. Table 10 reports Average Treatment Effects (ATE) with and without each mediator.

Table 10: Causal ablations on `OGBN-Arxiv` (batch size $B = 32$ vs. $B = 256$). Removing mediators reduces ATE magnitude, confirming their role in the causal pathway.

| Mediator Removed | ATE | % Drop from Full Model |
|---|---|---|
| None (Full Model) | +3.41 | — |
| Gradient Noise $N$ | +1.92 | −43.7% |
| Minima Sharpness $S$ | +2.18 | −36.0% |
| Model Complexity $C$ | +2.54 | −25.5% |


Figure 4: **Hessian Spectrum vs. Batch Size.** Largest Hessian eigenvalue $\lambda_{\max}$ increases with batch size $B$ across datasets (Cora, CiteSeer, PubMed, Amazon, OGBN-Arxiv), indicating sharper minima for larger $B$ and validating the trend $S(B) \propto 1/B$ (smaller $B \Rightarrow$ flatter).

### F.3. Alternative Causal Models

Replacing our hypergraph SCM with a linear DAG model (no higher-order edges) yields systematically weaker fits and ATE estimates (Table 11).

Table 11: Comparison of causal models for $B \to G$ estimation. HGCNet captures higher-order dependencies, improving ATE estimation stability.

| Model | ATE ($B = 32$ vs. $B = 256$) | Variance Across Runs |
|---|---|---|
| Linear DAG SCM | +2.04 | 0.37 |
| HGCNet SCM (ours) | **+3.41** | **0.14** |

### F.4. Geometric Validation: Hessian Spectrum

We computed the top-$k$ eigenvalues of the Hessian for varying $B$ using the Lanczos method (Yao et al., 2020). Figure 4 shows a clear inverse trend between $B$ and curvature, validating $S(B) \propto 1/B$.

### F.5. Do-Calculus Derivation for Causal Effects

To quantify the causal effect of batch size ($B$) on generalization ($G$), we compute:

$$P(G \mid do(B = b)) = \sum_{N,S,C} P(G \mid N, S, C) P(N, S, C \mid do(B = b)) \tag{18}$$

#### F.5.1. Average Treatment Effect (ATE) Estimation

We compute the Average Treatment Effect (ATE) for batch size $B$=16 vs. $B$=512

$$ATE(B = 16, B = 512) = \mathbb{E}[G \mid do(B = 16)] - \mathbb{E}[G \mid do(B = 512)] \tag{19}$$

Table 12: Average Treatment Effect (ATE) of batch size, comparing $B = 16$ vs. $B = 512$.

| **Dataset** | $B = 16$ | $B = 512$ |
|---|---|---|
| Cora | 83.9% ± 0.5 | 80.5% ± 0.7 |
| CiteSeer | 79.1% ± 0.4 | 76.0% ± 0.6 |
| PubMed | 88.2% ± 0.5 | 84.8% ± 0.7 |
| Amazon | 92.4% ± 0.5 | 89.0% ± 0.7 |
| OGBN | 73.2% ± 0.4 | 70.0% ± 0.6 |
| **ATE** ($B = 16$ **vs.** $B = 512$) | +2.6% | |

**Conclusion:** Smaller batch sizes significantly improve generalization by increasing gradient noise and promoting flatter minima.

## Appendix G. Extending to Hyperparameter Interactions

### G.1. Motivation

The main paper assumes causal sufficiency with mediators $\{N, S, \kappa, C\}$ linking batch size $B$ to generalisation $G$. However, batch size interacts closely with other hyperparameters, most notably *learning rate* ($L$), *optimiser choice* ($O$), and *weight decay* ($WD$). To relax the sufficiency assumption, we extend the hypergraph with these co-interventions, enabling estimation of *policy-conditional causal effects*.

### G.2. Extended Hypergraph and Structural Equations

We define an augmented vertex set

$$\mathcal{V}' = \{B, L, O, WD, N, S, \kappa, C, G\},$$

with new hyperedges

$$\{B, L, O\} \to N, \qquad \{S, WD\} \to C.$$

This captures the fact that gradient noise is shaped jointly by batch size, learning rate, and optimiser preconditioning, while sharpness-to-complexity interactions are modulated by weight decay.

The extended structural equations are

$$N = f_N(B, L, O, \epsilon_N), \tag{20}$$
$$S = f_S(N, \epsilon_S), \tag{21}$$
$$\kappa = f_\kappa(B, L, O, \epsilon_\kappa), \tag{22}$$
$$C = f_C(S, WD, \kappa, \epsilon_C), \tag{23}$$
$$G = f_G(C, \kappa, \epsilon_G). \tag{24}$$

### G.3. Policy-Conditional Causal Effects

We now study the policy-conditional interventional distribution

$$P(G \mid do(B = b), do(H \sim \pi_H)), \qquad H = \{L, O, WD\},$$

where $\pi_H$ denotes a hyperparameter policy (e.g. fixed learning rate, cosine decay, SGD vs. Adam). The *policy-conditional average treatment effect* (PC-ATE) is

$$\mathrm{ATE}_{b_1,b_2|\pi_H} = E[G \mid do(B = b_1), do(H \sim \pi_H)] - E[G \mid do(B = b_2), do(H \sim \pi_H)].$$

### G.4. Estimation Strategy

We extend the neural ridge surrogate to regress $G$ on $(M, H)$, where $M = \{N, S, \kappa, C\}$. To mitigate mediator–hyperparameter collinearity, we adopt an orthogonalised regression: first residualising mediators on $H$, then regressing $G$ on the residuals and batch size. This yields stable PC-ATE estimates under different training protocols.

### G.5. Illustrative Results

We conduct a $2 \times 2 \times 2$ factorial study on PubMed and OGBN-Arxiv, varying batch size, learning rate, and weight decay under SGD and Adam. Results confirm that smaller batches consistently improve generalisation, though effect magnitude depends on hyperparameter policy $\pi_H$.

Table 13: Policy-conditional ATE (%) between $B = 32$ and $B = 256$ under different hyperparameter settings. Results averaged over 5 runs.

| Policy $\pi_H$ | PubMed $\Delta G$ | OGBN-Arxiv $\Delta G$ |
|---|---|---|
| SGD, $L = 10^{-3}$, $WD = 0$ | +2.8 | +2.4 |
| SGD, $L = 3\times10^{-4}$, $WD = 5\times10^{-4}$ | +2.5 | +2.1 |
| Adam, $L = 10^{-3}$, $WD = 0$ | +2.3 | +1.9 |
| Adam, $L = 3\times10^{-4}$, $WD = 5\times10^{-4}$ | +2.1 | +1.7 |

### G.6. Key Observations

- The direction of the effect (smaller $B$ improves $G$) is invariant across $\pi_H$.
- The magnitude of improvement attenuates under Adam or lower learning rates, consistent with optimiser preconditioning dampening gradient noise.
- Weight decay reduces sharpness but also decreases the curvature-mediated gain, showing its dual role in complexity control.

## Appendix H. Statistical Significance of Batch Size Effects

We assess whether the accuracy gains from smaller batches are statistically significant by comparing $B = 16$ and $B = 512$ using 10 independent training runs per setting, each with a distinct random seed to account for variation from initialization, data shuffling, and optimization dynamics.

For each dataset, we compute the mean accuracy difference (16–512) and evaluate significance using a paired $t$-test and a Wilcoxon signed-rank test.

Table 14: Mean accuracy improvement of $B = 16$ over $B = 512$ with corresponding $p$-values. All values are averaged over 10 seeds; ± denotes standard error.

| **Dataset** | **Accuracy Gain (%)** | **$p$-value (t-test)** | **$p$-value (Wilcoxon)** |
|---|---|---|---|
| Cora | $+2.4 \pm 0.3$ | $3.1 \times 10^{-3}$ | $5.4 \times 10^{-3}$ |
| CiteSeer | $+2.1 \pm 0.3$ | $4.5 \times 10^{-3}$ | $7.8 \times 10^{-3}$ |
| PubMed | $+3.1 \pm 0.4$ | $1.2 \times 10^{-3}$ | $2.9 \times 10^{-3}$ |
| Amazon | $+3.2 \pm 0.4$ | $8.0 \times 10^{-4}$ | $1.5 \times 10^{-3}$ |
| OGBN-Arxiv | $+3.5 \pm 0.3$ | $5.0 \times 10^{-4}$ | $1.0 \times 10^{-3}$ |

**Observation:** All $p$-values are $< 0.01$, confirming that the performance gains from smaller batches are statistically significant across all datasets. This result reinforces our causal claim that reduced batch size enhances generalisation.

## Appendix I. Limitations

Our study focuses on isolating the causal role of batch size through a controlled hypergraph framework. This design necessarily introduces some limitations. First, we assume causal sufficiency with mediators $\{N, S, \kappa, C\}$, while holding other hyperparameters fixed; Appendix G shows how the framework can naturally extend to include learning rate, optimiser, and weight decay. Second, we estimate mediator effects using neural ridge regression. This choice prioritises stability under collinearity, though more expressive nonlinear estimators could further enrich the analysis. Third, curvature is measured via Ollivier–Ricci averages, which trade off fine-grained geometric detail for tractability. Finally, our experiments are restricted to representative citation graphs and text benchmarks; broader domains such as vision and reinforcement learning remain future work. These constraints reflect deliberate scope rather than fundamental barriers. The framework is general and could incorporate richer estimators, alternative curvature metrics, and multi-hyperparameter interventions, providing a pathway for extending causal analysis of training dynamics at larger scales.

## Appendix J. Impact Statement

This work provides a causal theoretic understanding of how batch size affects generalisation in deep learning. By isolating stochastic gradient noise as the primary driver of batch size induced variability, and validating this through do calculus interventions, the findings clarify long standing questions around training dynamics, convergence, and minima geometry. The discovery that hypergraph based causal models outperform pairwise approaches further advances structural learning methods, particularly in high dimensional or multiagent settings. These insights can inform the design of more robust and generalisable AI models across application domains such as healthcare, finance, and autonomous decision making. For instance, training protocols that account for the implicit regularisation effects of batch size may lead to improved diagnostic reliability in medical models, or more stable behaviour in financial forecasting systems. At the same time, faster convergence with large batches, if poorly understood, could lead to brittle deployments or unintended consequences, especially in high stakes environments. This research is primarily methodological, but its implications extend to practical and ethical considerations in model development. We recommend future work on adaptive training regimes that adjust batch size in response to curvature aware signals, and on integrating causal diagnostics into real world ML pipelines. As optimisation behaviour becomes increasingly consequential for model reliability and safety, rigorous analysis of training dynamics, such as that offered here, becomes ever more vital.