

Fast Linear Solvers via AI-Tuned Markov Chain Monte Carlo-based Matrix Inversion

Anton Lebedev*

Won Kyung Lee*

anton.lebedev@stfc.ac.uk

wonkyung.lee@stfc.ac.uk

STFC Hartree Centre

Warrington, United Kingdom

Vassilis Kalantzis

IBM Research

Yorktown Heights, New York, USA

vkal@ibm.com

Shashanka Ubaru

IBM Research

Yorktown Heights, New York, USA

shashanka.ubaru@ibm.com

Soumyadip Ghosh

IBM Research

Yorktown Heights, New York, USA

ghosh.s@us.ibm.com

Olha I. Yaman

STFC Hartree Centre

Warrington, United Kingdom

olha.ivanyshyn-yaman@stfc.ac.uk

Yingdong Lu

IBM Research

Yorktown Heights, New York, USA

yingdong@us.ibm.com

Lior Horesh

IBM Research

Yorktown Heights, New York, USA

lhoresh@us.ibm.com

Tomasz Nowicki

IBM Research

Yorktown Heights, New York, USA

tnowicki@us.ibm.com

Vassil Alexandrov

STFC Hartree Centre

Warrington, United Kingdom

vassil.alexandrov@stfc.ac.uk

Abstract

Large, sparse linear systems are pervasive in modern science and engineering, and Krylov subspace solvers are an established means of solving them. Yet convergence can be slow for ill-conditioned matrices, so practical deployments usually require preconditioners. Markov chain Monte Carlo (MCMC)-based matrix inversion can generate such preconditioners and accelerate Krylov iterations, but its effectiveness depends on parameters whose optima vary across matrices; manual or grid search is costly. We present an AI-driven framework recommending MCMC parameters for a given linear system. A graph neural surrogate predicts preconditioning speed from A and MCMC parameters. A Bayesian acquisition function then chooses the parameter sets most likely to minimise iterations. On a previously unseen ill-conditioned system, the framework achieves better preconditioning with 50% of the search budget of conventional methods, yielding about a 10% reduction in iterations to convergence. These results suggest a route for incorporating MCMC-based preconditioners into large-scale systems.

CCS Concepts

- Computing methodologies → Massively parallel and high-performance simulations; Linear algebra algorithms; Neural networks;
- Theory of computation → Numeric approximation algorithms.

Keywords

Markov Chain Monte Carlo, MCMCMI, Numerical Linear Algebra, AI, Recommendation Systems

Author's version. ©2025 The Authors. Publication rights licensed to ACM. One or more authors of this work are employees or contractors of a national government. This is the author's version of the work. It is posted here for your personal use; not for redistribution. The definitive Version of Record

*Both authors contributed equally to this research.

was published in *SC Workshops '25, November 16–21, 2025, St Louis, MO, USA*, <https://doi.org/10.1145/3731599.3767543>.

1 Introduction

Large and sparse systems of linear equations arise daily in computational fluid dynamics, structural analysis, plasma physics and countless other branches of modern science and engineering. Krylov subspace methods such as Conjugate Gradient (CG), Generalized Minimal RESidual method (GMRES) method, and their variants remain the work-horses for solving such systems because they require only matrix–vector products and a modest memory footprint [26]. Nevertheless, their convergence can be quite slow when the matrix is ill-conditioned, which is the case when the matrix represents a differential operator discretised on a fine mesh. In such cases, practical applications rely on preconditioners that improve the spectral properties of the system.

Classical algebraic preconditioners such as Incomplete LU (ILU) and Incomplete Cholesky (IC) factorisations are powerful yet may fail or become prohibitively expensive for highly irregular matrices. In response, Markov-chain Monte-Carlo (MCMC) matrix-inversion (MI) techniques have resurfaced as an appealing alternative preserving the sparsity of the matrix while offering a high degree of embarrassing parallelism. Despite these merits, the performance of MCMC preconditioning is acutely sensitive to the choice of hyperparameters, and optimal values generally differ from one linear system to the next. Exhaustive or grid searches over the parameter space require many expensive solver runs, thus undermining the practical advantage that MCMC was meant to provide.

In this paper, we tackle the bottleneck of algorithmic parameter tuning for MCMC preconditioners. In particular, we propose an AI-assisted framework that intelligently selects MCMC parameters for any given sparse system. At its core, our framework leverages a graph neural surrogate model. Given an iteration matrix A and

a set of candidate MCMC parameters, our model predicts the expected reduction in the number of iterations to achieve convergence compared to the unpreconditioned case. A Bayesian acquisition function then exploits the surrogate’s statistical nature, in each batch, those parameter sets which are most likely to minimise the expected iteration count. Applied to previously unseen matrices, the framework exceeds baseline preconditioning performance with only 50% of the search budget demanded by conventional methods and delivers about 10% fewer number of steps to convergence. These results indicate a practical route towards deploying MCMC preconditioners in large-scale linear-solver pipelines with minimal manual intervention.

The remainder of the paper is organised as follows. Section 2 reviews related work, covering both advances in preconditioning and MCMC-based MI. Section 3 introduces our proposed framework, including details of the surrogate model architecture and Bayesian selection loop. Section 4 details the experimental set up, while Section 5 discusses results and limitations. Section 6 concludes with avenues for future research.

2 Literature Review

Early work on algebraic preconditioning concentrated on deterministic approximations of A^{-1} . Incomplete factorisations [25] such as ILU and IC remain a staple of various large-scale scientific and engineering simulations. However, they are difficult to pipeline on modern hardware while ILU may break down for indefinite matrices [8, 35]. Sparse Approximate Inverse (SPAI) schemes address this parallelism bottleneck [10], constructing an explicit sparse stand-in for A^{-1} that can be applied via Sparse Matrix-Vector multiplications (SpMV)—an operation that parallelises well. Stochastic methods based on MCMC sidestep the triangular-solve bottleneck of factorisation-type preconditioners by estimating columns of A^{-1} through independent random walks [30]. Because each walk evolves independently, the work decomposes into parallel tasks that can scale almost linearly on modern accelerator hardware [16]. The approach has proved viable on large real-world test cases, including electromagnetic, climate, plasma physics simulations [27], while ongoing algorithmic advances—most recently the regenerative formulation that collapses multiple hyperparameters into a single transition budget parameter—continue to improve robustness and variance control [9]. Despite their architectural advantages, these methods have a critical practical flaw. Their effectiveness hinges on a set of hyperparameters whose optimal values are strongly matrix-dependent. Finding them often requires costly, manual trial-and-error, creating a significant tuning bottleneck.

On the other hand, there have been recent attempts to develop learning-based matrix preconditioners that aim to infer a good approximation of the inverse from data. In particular, [6] suggested a graph neural preconditioner that generates a preconditioner for a matrix in a linear system through a black box graph neural network model. Another significant area of interest is the specialisation of learning-based preconditioners for systems derived from Partial Differential Equations (PDEs), typically by utilising additional geometric or physical information from the underlying physical problem as additional machine learning model inputs. For instance, some methods learn from the problem’s geometry by using the

physical grid coordinates as model inputs [20, 31], whilst others learn from its physical by directly incorporating the PDE’s coefficients [18]. Such PDE-tailored methods are further constrained by their application-specific design. More fundamentally, learning-based matrix preconditioners suffer from inherit limitations related to their “black-box” nature, such as the difficulty in diagnosing and remedying poor performance on a specific matrix. This hinders their practical adoption, especially for mission-critical applications where reliability, stability and explainability are of paramount importance.

3 Methodology

In this paper, we consider the linear system $Ax = b$, where $A \in \mathcal{A} \subset \mathbb{R}^{n \times n}$. Our aim is to obtain a fast approximation of the inverse A^{-1} so that the system can be solved more efficiently. As a representative example, we adopt the MCMC-based MI schemes of the prior literature [16, 27]; more recent variants such as [9] could be also employed. An MCMC-based MI method requires a vector of algorithmic parameters $x_M \in \mathcal{X}_M$ and returns a preconditioner $P \approx A^{-1}$. We then solve $PAx = Pb$, where the Krylov solver will typically converge faster due to the lower condition number of PA .

Here we aim to identify parameters x_M of the MCMCMI method that will minimise the overall time-to-solution of the system, the time required to create the preconditioner and solve the system, accounting for matrix dependence via matrix A and its features $x_A \in \mathcal{X}_A$. To this end, we construct a surrogate model that, given (A, x_A, x_M) , predicts the resulting expected preconditioning speed $\mu(A, x_M)$. Specifically, the surrogate model outputs the predicted mean $\hat{\mu}$ together with an uncertainty estimate $\hat{\sigma}$. The optimal MCMC parameters are then defined by $x_M^*(A) = \arg \min_{x_M} \mu(A, x_M)$, and are selected via an acquisition function that balances exploration of the parameter space with exploitation of the surrogate’s current best estimate.

3.1 Graph Neural Surrogate Model

We employ a graph neural network f_θ as the surrogate model because its message passing operations are size-invariant so that the model can take varying sizes of matrices as parts of inputs. We design the graph neural surrogate model that extracts information directly from the matrix A and augments it with inexpensive matrix features x_A and the candidate MCMC parameter vector x_M . For this, we construct a weighted and directed graph $G = (V, x_V, E, w_E)$ from the matrix $A \in \mathbb{R}^{n \times n}$, whose vertex set $V = \{1, \dots, n\}$ represents the rows of A . An edge $(i, j) \in E$ exists iff $A_{ij} \neq 0$ and carries weight $w_E(i, j) = A_{ij}$. Each vertex stores the unweighted row degree $x_V(i) = \deg(i) = |\{j : A_{ij} \neq 0\}|$.

In addition to graph data, matrix dependence is also accounted for via matrix features $x_A \in \mathcal{X}_A$ that are cheap to compute, such as the norms, sparsity and symmetricity. All features are standardised—each value is rescaled to zero mean and unit variance—so that they contribute on a comparable scale during training.

The graph neural surrogate model receives the triplet (G, x_A, x_M) and processes each component separately before fusion. A stack of l_g message passing layers extracts a graph embedding h_g from G [1]. Although numerous message-passing formulations exist, most of them follow the similar basic pattern: at each layer, every node

aggregates information from its neighbours and itself, summarises (i.e., pools) these messages, and then applies a non-linear transformation to produce an updated representation. Several modern graph neural layer architectures are explored, and the most suitable one is selected through extensive experimentation. In parallel, l_A and l_M fully connected (FC) layers transform x_A and x_M into embeddings h_A and h_M . Layer normalisation is applied in both the message passing layers and FC stacks to stabilise training and mitigate covariate shift, while ReLU provides the non-linear activation. The three latent representations h_g , h_A , and h_M are then concatenated and passed through l_c FC layers with dropout, producing a vector h_{combined} . Finally, two linear heads provide the predicted mean and standard deviation of the MCMC preconditioning performance metric:

$$\hat{\mu} = \text{ReLU}(W_\mu h_{\text{combined}} + b_\mu), \hat{\sigma} = \ln(1 + e^{W_\sigma h_{\text{combined}} + b_\sigma}), \quad (1)$$

where W_μ , b_μ , W_σ and b_σ are trainable weights and biases. The former expression applies a ReLU to obtain the predicted mean, while the latter employs the soft-plus transform $\ln(1 + e^z)$ to ensure a strictly positive standard deviation. We model the MCMC preconditioning performance as a Gaussian distribution with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$.

Given a training dataset $\mathcal{D} = \{(G_i, x_{A,i}, x_{M,i}, \bar{y}_i, s_i)\}_{i=1}^N$, where \bar{y}_i and s_i are the sample mean and sample standard deviation of the repeated solver runs for the i -th input, we learn the surrogate model parameters by minimising a Mean Squared Error objective

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N [(\hat{\mu}_i - \bar{y}_i)^2 + (\hat{\sigma}_i - s_i)^2]. \quad (2)$$

A Gaussian negative log-likelihood could be also considered as an alternative, but very small s_i values could make that objective numerically unstable.

3.2 Acquisition Function

In most cases, we never get enough evaluations to exhaustively scan the MCMC parameter space. We therefore rely on Bayesian Optimisation (BO) to determine the next parameter vectors x_M to test. The surrogate model f_θ provides, for each candidate, a predictive mean $\hat{\mu}$ and a predictive uncertainty $\hat{\sigma}$ as in 1. An acquisition function balances two objectives: it exploits regions in which the predicted mean $\hat{\mu}$ is already low, while it simultaneously explores regions with high predictive uncertainty $\hat{\sigma}$, as those areas may still conceal better solutions. Past observations on related matrices, for example, smaller matrices representing the same differential operator, allow the surrogate model to transfer knowledge, thereby reducing uncertainty for similar systems.

While a variety of acquisition functions are available, we adopt Expected Improvement (EI) because it has been shown to deliver consistently lower simple regret than others, for instance, confidence-bound methods [29] across a large suite of benchmarks [21]. EI naturally balances exploitation and exploration while relying on a single and intuitive exploration parameter ξ [22, 23]. Setting $\xi = 0$ yields pure exploitation, whereas values in the range $0.01 - 0.10$ gradually favour uncertain regions of the search space. Moreover, for a Gaussian surrogate posterior EI has the closed form

$$\text{EI}(x_M) = (y_{\min} - \hat{\mu} - \xi) \Phi\left(\frac{y_{\min} - \hat{\mu} - \xi}{\hat{\sigma}}\right) + \hat{\sigma} \varphi\left(\frac{y_{\min} - \hat{\mu} - \xi}{\hat{\sigma}}\right), \quad (3)$$

Algorithm 1 Bayesian tuning loop for MCMC parameter selection

```

Require: evaluated matrix set  $\mathcal{A}_{\text{train}}$ , total budget  $B$ , batch size  $k$ 
    Initialise  $\mathcal{D}_0$  with coarse grid-search records  $(A, x_M, \bar{y}, s)$ 
    for  $t = 0, 1, \dots$  until  $|\mathcal{D}_t| = B$  do
        Fit surrogate  $f_\theta$  on  $\mathcal{D}_t$ 
        for all  $A \in \mathcal{A}_{\text{train}}$  do
            for  $j = 1$  to  $k$  do
                draw initial  $x_M^{(j,\text{init})}$ 
                 $x_M^{(j)} \leftarrow \text{L-BFGS-B}$  maximise  $\text{EI}(x_M; A)$  starting from  $x_M^{(j,\text{init})}$ 
                Run MCMC + Krylov solver (e.g., GMRES) with  $x_M^{(j)}$ 
                Record  $(A, x_M^{(j)}, \bar{y}, s)$  and append to  $\mathcal{D}_{t+1}$ 
            end for
        end for
    end for
    return  $x_M^\star(A) = \arg \max_{x_M} \text{EI}(x_M; A)$  given  $A \in \mathcal{A}$ 

```

where y_{\min} is the best MCMC preconditioning performance metric observed so far and Φ and φ are the standard normal Cumulative Distribution Function and Probability Density Function, respectively. The first term measures the expected drop below the current best exploitation, whereas the second term rewards large predictive variance exploration. Because EI is differentiable with respect x_M , we can maximise it efficiently with first-order optimisers. In practice, we minimise the negative EI using the gradient-based quasi-Newton method L-BFGS-B [5]. At every step the candidate x_M is fed through the surrogate model; back-propagation supplies the exact gradient $\nabla_{x_M} [-\text{EI}(x_M)]$, which L-BFGS-B exploits to build curvature information and update the iterate. Algorithm 1 summarises the complete optimisation loop.

4 Experiments

4.1 MCMC Preconditioning

To benchmark our tuning framework we adopt the advanced MCMC-based MI preconditioner of [16, 27]. The method is governed by three continuous algorithmic parameters $x_M = (\alpha, \varepsilon, \delta)$:

- $\alpha \in \mathbb{R}_{>0}$: a matrix perturbation parameter to scale the added diagonal of A so that the Neumann-series preconditioner converges;
- $\varepsilon \in (0, 1]$: a stochastic error that determines the maximum number of independent Markov chains;
- $\delta \in (0, 1]$: a truncation error that determines the maximum walk length of a Markov chain.

In addition, x_M includes a categorical variable for the Krylov solver type. Due to the limited size of our dataset, we do not attempt to provide a recommendation of the solver type here. Two matrix-independent settings are fixed: the filling factor, which controls the number of non-zeros retained in the preconditioner, and the truncation threshold. The preconditioner's filling factor is set to $2\phi(A)$, where $\phi(A)$ is that of the original matrix A , whereas the truncation threshold is set arbitrarily to 10^{-9} to avoid introducing truncation of the preconditioners. Also, all preconditioners are obtained using a hybrid MPI+OpenMP code running on a single node utilising 2 MPI processes with 4 threads per process. The preconditioned system is then solved with GMRES or BiCGStab; when the matrix A is symmetric positive definite, we also employ

CG. Also, we define the MCMC preconditioning performance metric

$$y(A, x_M) = \frac{\# \text{ of steps with preconditioner}}{\# \text{ of steps without preconditioner}}, \quad (4)$$

so that the optimiser seeks the x_M minimising this ratio for every matrix.

4.2 Dataset

We built the dataset from 11 sparse matrices summarised in Table 1. Each entry appears with its dimension n , symmetricity, and condition number $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$. The selection covers the archetypal 2D Finite-Difference (FD) Laplacian matrix as a representative of FD methods and symmetric positive-definite matrices, a0XXXX matrices, representing asymmetric differential operators from plasma physics discretised using finite elements at various mesh resolutions, a finite-element discretisation of an unsteady advection-diffusion problem with varying mesh resolutions, PDD_RealSparse, and a representative of systems occurring in climate simulations (nonsym_r3_a11). In FD or finite-element discretisations on shape-regular meshes, the condition number of the matrix for an m -th order PDE operator scales with the mesh width h as $O(h^{-m})$. In particular, $O(h^{-2})$ scaling is illustrated in Table 1 for the 2D FD Laplacian matrix. A large condition number can severely degrade the performance of iterative solvers, making effective preconditioning essential for maintaining computational efficiency.

To obtain the basis dataset for training we used a $4 \times 4 \times 4$ grid of parameters $\alpha \in \{1, 2, 4, 5\}$, $\epsilon \in \{1/2, 1/4, 1/8, 1/16\}$, $\delta \in \{1/2, 1/4, 1/8, 1/16\}$ with each α, ϵ, δ configuration executed ten times with GMRES and BiCGStab. The resulting sample mean and standard deviation of the performance metric $y(A, x_M)$ constitute one labelled datum per solver. Hence every matrix contributed 64 samples per solver (128 in total for the two-solver case). The symmetric Laplace matrices were additionally run with CG at $\alpha = 0.1$. A few samples with near-zero α were added to expose the surrogate to divergence scenarios. In total, the dataset for training and validation contains 1,318 labelled points, which were split 80%/20% into training and validation sets.

On the other hand, generalisability was assessed on the higher-order unsteady_adv_diff_order2_0001 ($\kappa \approx 6.6 \times 10^6$), a substantially harder system than its order-1 counterpart in the training phase; success here demonstrates that information transfers to an unseen ill-conditioned system.

4.3 Hyperparameters

We performed extensive hyperparameter tuning for both the graph neural surrogate model and the acquisition function to maximise predictive accuracy and optimisation efficacy. For the graph convolutional architecture, we considered six representative message-passing mechanisms: GATv2[3], Graph Transformer[28], GMM-Conv [24], EdgeConv[32], GINE[11], and PNA[7]. These were combined with three neighbourhood aggregation strategies: MultiAggregation [7], MeanAggregation, and DeepSetsAggregation[4]. We searched over hidden dimensions {32, 64, 128, 256, 512} and up to four message-passing layers. For the auxiliary inputs x_A and x_M , FC layers were employed with one to four layers, and hidden dimensions chosen from {8, 16, 32, 64} for x_A and {4, 8, 16, 32} for

Table 1: Matrix set used for this study

Matrix	Dimension	Symmetricity	$\kappa(A)$	$\phi(A)$
2DFDLaplace_16	225	Yes	1.0×10^2	0.042
2DFDLaplace_32	961	Yes	4.1×10^2	0.001
2DFDLaplace_64	3,969	Yes	1.7×10^3	0.0024
2DFDLaplace_128	16,129	Yes	6.6×10^3	0.0006
nonsym_r3_a11	20,930	No	1.9×10^4	0.0044
a00512	512	No	1.9×10^3	0.059
a08192	8,192	No	3.2×10^5	0.0007
unsteady_adv_diff_order1_0001	225	No	4.1×10^6	0.646
unsteady_adv_diff_order2_0001	225	No	6.6×10^6	0.646
PDD_RealSparse_N64	64	No	1.3×10^1	0.1
PDD_RealSparse_N128	128	No	5.0	0.1
PDD_RealSparse_N256	256	No	7.0	0.1

x_M . The concatenated embedding was passed through another FC block with hidden dimensions in {32, 64, 128, 256, 512} and up to four layers. We fixed the batch size at 128.

Continuous hyperparameters included the learning rate, sampled from a log-uniform distribution between 10^{-4} and 10^{-1} , the weight decay parameter from 10^{-6} to 10^{-3} , and dropout rates uniformly from 0 to 0.2. Hyperparameter optimisation was performed using the Tree-structured Parzen Estimator [2]. We used the Asynchronous Successive Halving Algorithm scheduler [17] for early stopping and resource-efficient scheduling, with a maximum of 150 epochs, a grace period of 20, and a reduction factor of 3. A total of 30 trials were launched, each corresponding to a different model configuration. In the acquisition function, we tested both a balanced strategy with $\xi = 0.05$ and an exploration-heavy strategy with $\xi = 1.0$.

4.4 Experimental Results

We assess the proposed pipeline on an unseen, highly ill-conditioned testing matrix. The graph neural surrogate was trained by a single NVIDIA V100 GPU (32 GB), and all other experiments were executed on CPUs.

The surrogate trained on the training dataset is referred to as PRE-BO MODEL. The best graph neural surrogate model was selected through the Bayesian Optimisation-based hyperparameter optimisation (HPO), based on the average performance across three random seeds. The selected surrogate model architecture include learning rate as 1.848×10^{-3} , weight decay parameter as 1, a single Edge Convolutional layer [32] and mean neighbourhood aggregation function with 256 hidden dimension for graph embedding, while a single FC layer with 64 dimension for embedding of the matrix feature x_A and three FC layers with 16 hidden dimension for MCMC algorithmic parameter x_M embedding. Also, two FC layers were added to represent the combined representation with 128 hidden dimension. The Adam optimiser [15] was used for model training. Although model training and model selection with HPO required approximately seven hours of compute time, such cost is expected to be amortised when applying the framework to large-scale matrices, where reduced solution cost can yield substantial overall savings.

To assess how the GNN surrogate evolves when augmented with new, targeted data, we performed one round of BO using the EI acquisition function (3) with two settings: $\xi = 0.05$ (balanced search) and $\xi = 1.00$ (exploration search). The PRE-BO MODEL was used to recommend a batch of 32 candidate x_M vectors for each BO strategy, for which MCMC preconditioning metrics were measured (10 replicates each). These new measurements were combined with the original training dataset to retrain the surrogate, producing the BO-ENHANCED MODEL. During retraining, we reused the hyperparameters selected for the PRE-BO MODEL and re-optimised only the model weights. Model retraining required approximately an hour of compute time. Performance was evaluated on an unseen, ill-conditioned testing matrix `unsteady_adv_diff_order2_0001`, using experimental results from a grid search over 64 distinct x_M vectors (10 replications each, 640 observations in total).

First, we assessed the reliability of the surrogate models' uncertainty estimates, whether the predicted confidence intervals capture the observed variability at the expected rate. This calibration assessment reveals whether the model is overconfident (intervals too narrow) or underconfident (intervals too wide), which is critical when the surrogate model is used to guide BO decisions.

Figure 1 shows calibration curves comparing the expected proportion of observations within the predicted interval (x-axis) to the actual proportion observed (y-axis) for the corresponding confidence levels $\tau \in \{0.50, 0.68, 0.80, 0.90, 0.95, 0.99\}$. For each τ , the symmetric prediction interval was defined as

$$[\hat{\mu}_j - z_{(1+\tau)/2} \hat{\sigma}_j, \hat{\mu}_j + z_{(1+\tau)/2} \hat{\sigma}_j], \quad (5)$$

where $(\hat{\mu}_j, \hat{\sigma}_j)$ are the surrogate model's predicted mean and standard deviation, and j indexes the 640 individual observations (64 distinct x_M , each with 10 replicates), with $(\hat{\mu}_j, \hat{\sigma}_j)$ identical within replicates of the same x_M . The empirical coverage \hat{p} was computed as the proportion of observations y_j falling within this interval.

To quantify the sampling uncertainty in \hat{p} , we computed the two-sided Wilson score 95% confidence interval for a binomial proportion [34]:

$$\text{CI}_{\text{Wilson}}(\hat{p}) = \frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \quad (6)$$

where n is the number of observations and $z = z_{0.975}$. This method is preferred over the normal approximation because it produces well-behaved bounds in $[0, 1]$, even for small n or extreme proportions. Shaded bands in Figure 1 represent these Wilson intervals. According to the plot, the PRE-BO MODEL exhibits clear under-coverage (overconfidence), with curves lying below the ideal diagonal. After a single BO round, the BO-ENHANCED MODEL shifts markedly closer to the diagonal, indicating improved calibration. In particular, for higher α values ($\alpha = 4.0$ and $\alpha = 5.0$), coverage approaches the ideal line, and the improvement is statistically significant according to the Wilson intervals (6). On the other hand, when the iteration matrix is ill-conditioned, lower values of α generally result to similar measurements, which in turn limits the learning ability of the surrogate model.

Further we examined whether the surrogate model's predicted mean lies within the empirical confidence interval for each set of MCMC algorithmic parameters x_M . Specifically, for each of the

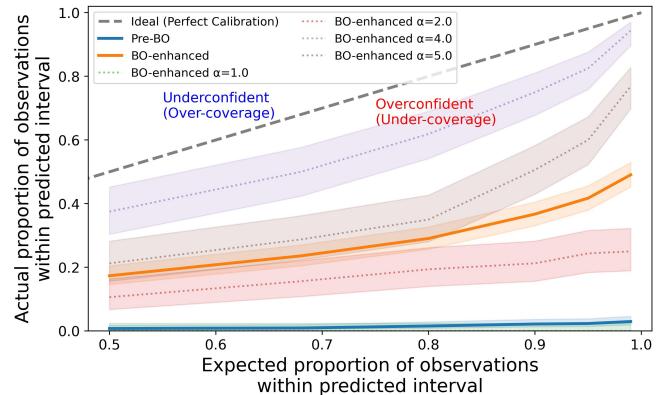


Figure 1: Calibration plot comparing predicted and observed coverage probabilities across multiple confidence levels

64 distinct x_M , we computed the sample mean \bar{y} , standard deviation s and 99% confidence interval of the preconditioning performance metric $y(A, x_M)$ across 10 replications. We then investigated whether the model's predicted mean falls inside this empirical interval. This pointwise coverage analysis focuses on the accuracy of the predicted central value at each algorithmic parameter point, rather than on the coverage of the observed data by the model's own predicted intervals. According to the plot shown in Figure 2, the PRE-BO MODEL (top row) does not represent reality well, as its mean frequently lies outside of the confidence interval. In contrast, the BO-ENHANCED MODEL (bottom row) achieves substantially higher inclusion over broad regions across the (ε, δ) grid for higher alphas $\alpha \in \{4.0, 5.0\}$. The heatmaps show, that contrary to prior assumptions [16], ε and δ do not contribute symmetrically to the success of the preconditioner. We observe that, given a truncation error δ , a successful preconditioner is obtained if $\varepsilon \lesssim \delta$, with this condition being more pronounced at larger perturbations α . Conversely, for a fixed stochastic error ε , thus a fixed number of Markov chains, a larger δ and thus shorter chains are preferable. Since larger ε and δ will correspond to shorter preconditioner computation, we may conclude that, for a fixed α , there will be an optimal combination ε^*, δ^* in the vicinity of the diagonal $\varepsilon = \delta$ that will minimise the overall runtime to solution. We further observe that no notable reductions in solver steps are achieved for parameter combinations $\varepsilon, \delta \ll \varepsilon^* \approx \delta^*$.

Finally, we assessed the practical utility of algorithmic parameter search for MCMC preconditioning on the ill-conditioned testing matrix `unsteady_adv_diff_order2_0001`. Despite using only 50% of the evaluation budget (32 recommendations) compared to grid or random search (64 recommendations), the BO-enhanced recommendations reduced the number of steps to convergence through MCMC preconditioning by up to 25%, which is approximately 10% fewer steps than grid search (Figure 3). The box plot summarises the distribution of the sample medians from 10 replications over the explored candidates of algorithmic parameters x_M . In addition, the coloured circle points show the distribution of observations $y(A, x_M^*)$ over 10 replications, where x_M^* denotes the single best

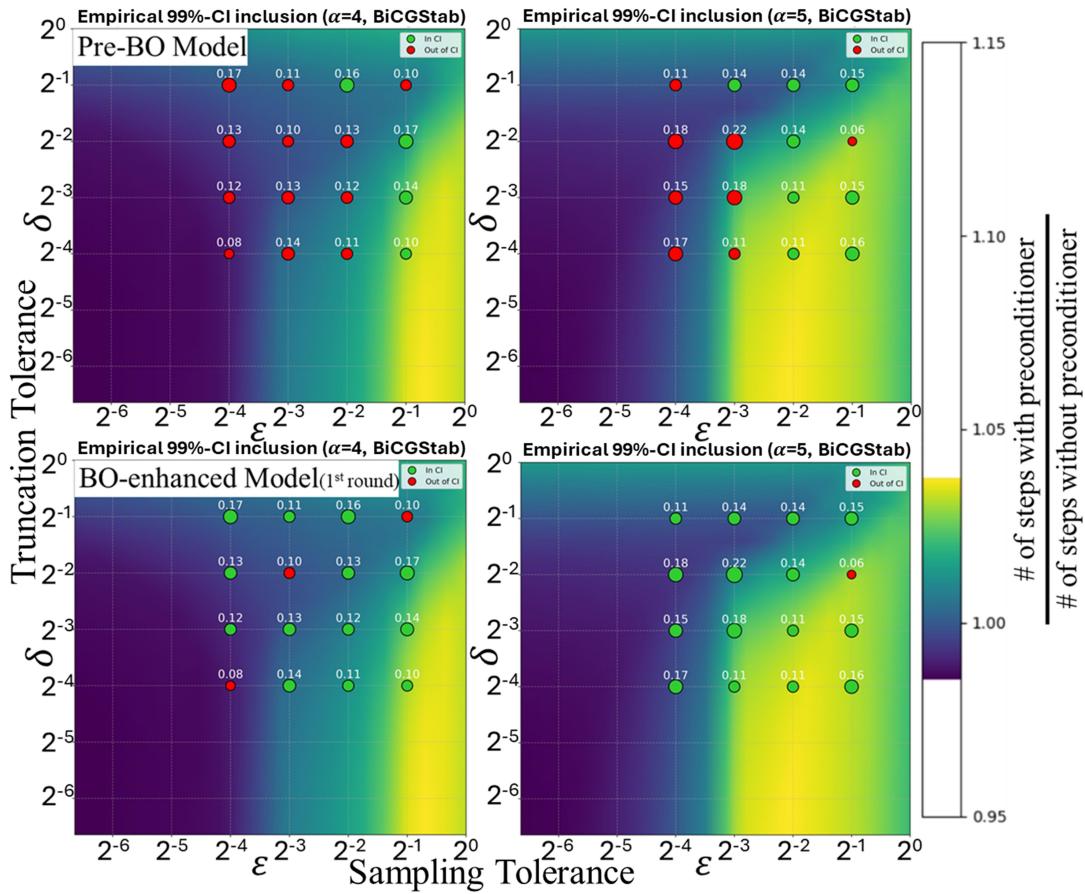


Figure 2: Confidence interval inclusion results showing the improved prediction accuracy of the surrogate model after BO retraining (BO-enhanced, bottom) compared to the baseline (Pre-BO, top)

recommendation of the algorithmic parameter that yields the minimum sample median of the MCMC preconditioning performance metric among all explored candidates for each search strategy.

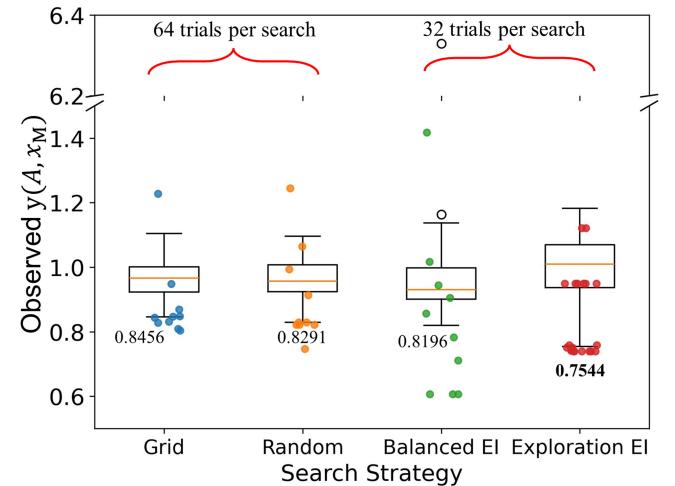


Figure 3: Box plot of sample median of $y(A, x_M)$ over the explored x_M , including the minimum. Coloured circle points represent the distribution of the observed $y(A, x_M^*)$ over 10 replications, where x_M^* indicates the parameter yielding the minimum sample median.

5 Conclusions

We have introduced a BO framework that couples a graph neural surrogate with an EI-based acquisition strategy to tune MCMC preconditioners for solving linear systems. On an ill-conditioned matrix unseen during the training phase, the method required only 50% of the search budget of a coarse grid yet reduced Krylov iterations by $\sim 10\%$. These results demonstrate that structural information extracted from the matrix A can guide hyper parameter search more efficiently than exhaustive sampling.

Several directions could further widen the scope and impact of the framework. First, although the Krylov method (GMRES, BiCGStab, CG) was included as a categorical input to the surrogate model, in this work we did not attempted to recommend the solver itself. Extending the framework to make such recommendations — selecting both the best solver and its optimal $(\alpha, \epsilon, \delta)$ given inexpensive matrix features such as symmetry, approximate condition number, and sparsity — would be a next step. Also, practical deployments must balance iteration speed-up against preconditioner build time; this could be achieved by adding parameters that govern the filling fraction and truncation threshold, whose costs scale linearly with non-zeros and retained elements, respectively, as well as hardware knobs such as thread count and MPI ranks. Such considerations can be especially challenging in substructuring eigenvalue solvers due to the implicit nature of the Schur complement matrix [12–14]. Also, future work could extend the framework to distributed-memory settings, explicitly minimising latency and accounting for communication and memory-management overheads to achieve robust scalability on parallel clusters.

Furthermore, the current FC layer-based forecasting in the last few layers of the graph neural surrogate model could be improved by replacing it with deep kernel learning [33] or a scalable Gaussian Process layer [19] for forecasting the MCMC preconditioning metric and its uncertainty. This, in turn, would enhance the quality of EI. In addition, consideration of reinforcement learning approaches that propose a set of correlated MCMC algorithmic parameter vectors could exploit structure in the search space more effectively than independent EI maximisation. Furthermore, an active learning loop or generative model that generates new linear systems for evaluation would allow the surrogate to evolve continually towards broader classes of matrices. On the other hand, the practical utility of the framework could be improved by employing a cost-sensitive, weighted loss that places greater weight on larger systems, thereby yielding better parameter recommendations for them. Finally, identifying when to switch from full retraining to fine-tuning of the surrogate model is an important direction for improving the framework’s practical applicability. Pursuing these directions will ultimately pave the way for more general and efficient linear-system solvers that accelerate scientific discovery across a broad range of applications such as climate modelling, computational fluid dynamics and plasma physics.

Acknowledgments

This work was supported by the Hartree National Centre for Digital Innovation, a UK Government-funded collaboration between STFC and IBM.

References

- [1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).
- [2] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems* 24 (2011).
- [3] Shaked Brody, Uri Alon, and Eran Yahav. 2022. How Attentive are Graph Attention Networks? In *International Conference on Learning Representations*. <https://openreview.net/forum?id=F72ximsx7C1>
- [4] David Buterez, Jon Paul Janet, Steven J Kiddie, Dino Ogle, and Pietro Liò. 2022. Graph neural networks with adaptive readouts. *Advances in Neural Information Processing Systems* 35 (2022), 19746–19758.
- [5] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16, 5 (1995), 1190–1208.
- [6] Jie Chen. 2025. Graph Neural Preconditioners for Iterative Solutions of Sparse Linear Systems. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=Tkkrm3pA35>
- [7] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. 2020. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems* 33 (2020), 13260–13271.
- [8] Geoffrey Dillon, Vassilis Kalantzis, Yuanzhe Xi, and Yousef Saad. 2018. A hierarchical low rank Schur complement preconditioner for indefinite linear systems. *SIAM Journal on Scientific Computing* 40, 4 (2018), A2234–A2252.
- [9] Soumyadip Ghosh, Lior Horesh, Vassilis Kalantzis, Yingdong Lu, and Tomasz Nowicki. 2025. Regenerative Ulam-von Neumann Algorithm: An Innovative Markov chain Monte Carlo Method for Matrix Inversion. *SIAM Journal on Matrix Analysis and Applications (to appear)* (2025).
- [10] Marcus J Grote and Thomas Huckle. 1997. Parallel preconditioning with sparse approximate inverses. *SIAM Journal on Scientific Computing* 18, 3 (1997), 838–853.
- [11] Weihua Hu*, Bowen Liu*, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2020. Strategies for Pre-training Graph Neural Networks. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HJWWJSFDH>
- [12] Vassilis Kalantzis. 2020. A domain decomposition Rayleigh–Ritz algorithm for symmetric generalized eigenvalue problems. *SIAM Journal on Scientific Computing* 42, 6 (2020), C410–C435.
- [13] Vassilis Kalantzis and Lior Horesh. 2023. Enhanced algebraic substructuring for symmetric generalized eigenvalue problems. *Numerical Linear Algebra with Applications* 30, 2 (2023), e2473.
- [14] Vassilis Kalantzis, Yuanzhe Xi, and Lior Horesh. 2021. Fast randomized non-Hermitian eigensolvers based on rational filtering and matrix partitioning. *SIAM Journal on Scientific Computing* 43, 5 (2021), S791–S815.
- [15] Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*. 1–15.
- [16] Anton Lebedev and Vassil Alexandrov. 2018. On advanced Monte Carlo methods for linear algebra on advanced accelerator architectures. In *2018 IEEE/ACM 9th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA)*. IEEE, 81–90.
- [17] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. 2020. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems* 2 (2020), 230–246.
- [18] Yichen Li, Peter Yichen Chen, Tao Du, and Wojciech Matusik. 2023. Learning preconditioners for conjugate gradient PDE solvers. In *International Conference on Machine Learning*. PMLR, 19425–19439.
- [19] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. 2020. When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems* 31, 11 (2020), 4405–4423.
- [20] Jian Luo, Jie Wang, Hong Wang, Zijie Geng, Hanzhu Chen, Yufei Kuang, et al. 2024. Neural Krylov iteration for accelerating linear system solving. *Advances in Neural Information Processing Systems* 37 (2024), 128636–128667.
- [21] Erich Merrill, Alan Fern, Xiaoli Fern, and Nima Dolatnia. 2021. An empirical study of bayesian optimization: Acquisition versus partition. *Journal of Machine Learning Research* 22, 4 (2021), 1–25.
- [22] Jonas Mockus. 1994. Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization* 4, 4 (1994), 347–365.
- [23] Jonas Mockus. 1998. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization* 2 (1998), 117.
- [24] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. 2017. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1921–1929.

- vision and pattern recognition*. 5115–5124.
- [25] Yousef Saad. 1994. ILUT: A dual threshold incomplete LU factorization. *Numerical linear algebra with applications* 1, 4 (1994), 387–402.
 - [26] Yousef Saad. 2003. *Iterative methods for sparse linear systems*. SIAM.
 - [27] Emre Sahin, Anton Lebedev, Maksims Abalenkovs, and Vassil Alexandrov. 2021. Usability of Markov chain Monte Carlo preconditioners in practical problems. In *2021 12th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA)*. IEEE, 44–49.
 - [28] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. 2021. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 1548–1554.
 - [29] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. 2009. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995* (2009).
 - [30] J Straßburg and Vassil N Alexandrov. 2013. A Monte Carlo approach to sparse approximate inverse matrix computations. *Procedia Computer Science* 18 (2013), 2307–2316.
 - [31] Vladislav Trifonov, Alexander Rudikov, Oleg Iliev, Yuri M Laevsky, Ivan Oseledets, and Ekaterina Muravleva. 2024. Learning from linear algebra: A graph neural network approach to preconditioner design for conjugate gradient solvers. *arXiv preprint arXiv:2405.15557* (2024).
 - [32] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)* 38, 5 (2019), 1–12.
 - [33] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. 2016. Deep kernel learning. In *Artificial intelligence and statistics*. PMLR, 370–378.
 - [34] Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212.
 - [35] Tianshi Xu, Vassilis Kalantzis, Ruipeng Li, Yuanzhe Xi, Geoffrey Dillon, and Yousef Saad. 2022. parGeMSLR: A parallel multilevel Schur complement low-rank preconditioning and solution package for general sparse matrices. *Parallel Comput.* 113 (2022), 102956.

accepted 5 September 2025